A Study of Features and Processes Towards Real-Time Speech Word Recognition.

Tracy M. Clark

A thesis presented for the degree of Doctor of Philosophy
in Electrical and Electronic Engineering at the
University of Canterbury, Christchurch, New Zealand.

May 1993

ABSTRACT

Word recognition techniques are reviewed. An exhaustive comparative study of many of the factors that affect recognition accuracy is presented. Experiments centred on four major areas of word recognition are described: pre-processing techniques, recognition features, recognition algorithms and distance measures. Recognition accuracy, in the context of each of these four areas, is investigated using the digit vocabulary spoken by 10 New Zealand (6 male and 4 female)and 38 American (20 male and 18 female) speakers. Pre- processing techniques examined are the type of window, the length of the data frame, data frame overlap, and pre-emphasis. Acoustic features tested include temporal features such as energy and zero-crossing rate, as well as frequency based acoustic representations such as linear prediction coefficients, cepstral coefficients, dynamic (transitional) cepstral coefficients, and perceptual linear prediction coefficients. Three types of distance measures are also reported on the Euclidean, the weighted Euclidean, and the projection. Two methods of training, random template selection and clustering, are investigated. Accuracy improvement by combining different features is also examined. Implementation of a real-time word recognition system designed on the basis of the comparative study and experiments, is described. The system is based on a TMS320C30 and takes around 0.03 seconds per recognition. The real-time system achieves speaker-dependent accuracies greater than 95% and speaker-independent accuracies greater than 70% for the digit vocabulary. An examination is also made of two methods of continuous recognition using sub-word representations. Both these methods take advantage of isolated word recognition techniques such as dynamic programming. A segmentation method and anon-segmentation method were investigated. Accuracy of the segmentation recognition method is found to depend linearly on the accuracy of the segmenter. With a segmentation error of 22%, an average recognition accuracy of 90.7% was obtained for 10 vowels and 2 consonants. For the non-segmentation recognition method, an average accuracy of 75% was obtained. Although the segmentation method produced higher accuracies than the non-segmentation method, it is argued that the removal of the segmentation is an advantage that greatly simplifies the recognition strategy.

ACKNOWLEDGEMENTS

Because it is impossible to undertake the challanges in life alone it has been my good fortune during the course of my research and thesis preparation to have received assistance, advice, support, and encouragement from many people.

In particular I would like to thank my supervisors, the late Professor Richard Bates, for his continual guidance during the first stages of my research, and the energetic Bill Kennedy for often having to set me straight on many points during the full term of my research. The help from Dr Kathy Garden has also been much appreciated.

The production of a thesis is not only based on research, much of my learning was during the writing of the text. I would like to thank Bill Kennedy for his untiring work in turning my sometime garbled English into coherent discussion. I would also like to thank all the proof-readers which included my mother, Evelyn Clark, without her help I may have never finished.

I would also like to acknowledge the help I received through the stimulating discussions held with the Speech Group members, both formally and informally, these discussions including Andrew Elder, Catherine Watson, William Thorpe, John Kirkland, and Lim Ching Aun.

Because I found the production of this thesis sometime extremely daunting and particularly hard going I would like to thank all those that supported me during that time. In particular I would like to thank all my family and friends which stayed with me through all my bad (and good) times, these include Stephen Burdon, Sonya Clark, Ken Clark, Tony Green, and Catherine Watson.

I would like to acknowledge the financial support from Unisys Linc which allowed me to continue my studies to a PhD level.

PREFACE

This thesis presents the design and implementation of a real-time speech recognition system. The design of the recognition system has involved a comprehensive comparison of many major techniques used in word/speech recognition.

I was introduced to speech processing in late 1986 when I joined the speech group at University of Canterbury. The speech group, led by Professor Richard Bates, consisted of my contemporaries William Thorpe, Andrew Elder and Catherine Watson together with members of staff Mr Bill Kennedy and Dr Kathy Garden. The projects undertaken by the speech group were, and continue to be, firmly based in the area of real-time speech processing, and encompassed research into the design and implementation of a real-time speech therapy aid for the disabled speaker, techniques for low data rate speech encoding and the development of speaker and speech recognition algorithms.

As an introduction into this area of research I began by extending the capabilities of the real-time speech therapy aid, a project which had been going for some years. The development of the therapy aid provided me with a good basis in the designing and programming of real-time software. The computer-based speech therapy tool operates on a PC in conjunction with a TMS32010 digital signal processor (DSP). Speech from a microphone is input into the DSP chip via an analogue to digital (A-to-D) converter and the sampled speech processed by the DSP. The DSP outputs parameters, such as energy, pitch, or vocal tract parameters, which are transferred to the PC for graphical representation. Plotting is achieved in real-time; that is, as the speaker is talking into the microphone pictures of the speaker's vocal tract, intensity or pitch are drawn on to the screen. My work in this project entailed the designing and programming of a high-frequency monitor (fricative monitor) and the programming of a real-time spectrogram. This project gave me a solid grounding in DSP technology and the design of real-time applications. It also afforded me an appreciation of signal processing techniques.

I was becoming increasingly interested in using aspects of the algorithms designed for the therapy tool as a basis for designing a real-time speech recognition algorithm. The experience I had obtained as a Masters student encouraged me to plan towards a PhD and in April 1988 I was lucky enough to obtain the additional funding I needed to continue my research as a PhD student. Initially it was my desire to build a recognition system that would aid the disabled speaker. However, in order to attract funding from commercial interests, the project became one of designing a real-time word recognizer trainable for any speaker.

My research project aim became one of designing and building a highly accurate recognition system within the limitation of real-time operation. My problem was therefore to find the best methods of achieving this goal. Hence an examination of the major variables involved in a system, bounded by the real-time limitations, were undertaken. These variables fall into four categories; the pre-processing of the data, the extraction of features characterising a word, the measure of the difference between two sets of features (the distance measures) and the classification algorithm. An attempt to understand these four major categories which determine the fundamental and practical

limitations of speech recognition algorithms constitute the major original contribution of the thesis.

An in depth comparative study of pre-processing techniques(pre-emphasis, windowing, frame overlap and frame size), feature extraction methods (energy, zero-crossing, LPC, cepstral and perceptual based LPC) and distance measures (Euclidean, weighted Euclidean, log likelihood and projection) were trialed. Initially the techniques were applied to New Zealand speech but this made it difficult to compare results with those reported in the literature. Nevertheless these tests are important because a prime requirement of the recognition scheme is that it work with New Zealand speakers. For the results to be useful on a wider scale, however, the same comparative tests were undertaken with an American speech data-base bought from National Institute of Standards and Technology (NIST).

The major conclusions from the tests undertaken and reported in Chapter 8 of this thesis affirm that highly accurate speaker dependent recognition is possible, and furthermore the optimum choice of variables is not highly sensitive to accent. The most beneficial (giving the highest accuracy increase) pre-processing technique was found to involve windowing the data with a smooth function such as a Hamming window. The second most significant pre-processing technique was to pre-emphasise the speech. The most time consuming and worthless operation was to overlap the data frames. The most useful techniques to speed up operation without significantly reducing accuracy were to incorporate a threshold in the warping algorithm, to increase frame size and to decrease reference template numbers. Examining the features tested show that those techniques that gave higher recognition accuracies for any individual speaker tended to also give higher accuracies for all speakers, regardless of accent. Recognition accuracies can be further increased by applying a distance measure which enhances the feature set chosen.

Publications and presentations prepared during the course of my PhD. research are listed below.

BATES, R.H.T., BRIESEMAN, N.P., CLARK, T.M., ELDER, A.G., FRIGHT, W.R., GARDEN, K.L., KENNEDY, W.K., SQUIRES, P.L., THORPE, C.W., TURNER, S.G. and JELINEK, H.J. (1987), 'Interactive speech-defect diagnostic/therapeutic /prosthetic aid', In LETELLIER, J.P. (Ed.), *Real Time Signal Processing X*, Proceedings of SPIE - The International Society for Optical Engineering, 20-21 August, Pp. 131–139.

ELDER, A., BATES, R., BRIESEMAN, N., CLARK, T., FRIGHT, W., GARDEN, K., KENNEDY, W., SQUIRES, P., TURNER, S. and THORPE, C. (1987), 'Real-time speech therapy aid', *Proceedings of the 24th National Electronics Conference*, Vol. 24, September, Pp. 115–118.

WATSON, C.I., CLARK, T.M., ELDER, A.G. and THORPE, C.W. (1988), 'Multifarious real-time speech processing applications', *Proc. NELCON, (New Zealand National Electronics Conference)*, Vol. 25, Pp. 65–70.

CLARK, T., KENNEDY, W. and BATES, R. (1989), 'Towards a real-time computer word recognition system using the tms32030', *Proc. NELCON, (New Zealand National Electronics Conference)*, Vol. 27, P. .

AUN, L.C., ELDER, A., CLARK, T. and BATES, R. (1990), 'Software implementation of hidden Markov model for recognition of isolated digits by New Zealand speaker', *Proc. NELCON, (New Zealand National Electronics Con-*

ference), Pp. 287–294.

CLARK, T., KENNEDY, W. and BATES, R. (1990), 'Features for a computer word recognition system', *Proceedings of the Third Australian international conference on Speech science and technology.*, November, Pp. 356–361.

CLARK, T., KENNEDY, W. and BATES, R. (1991), ' A real-time word recognition system', *Proc. NELCON, (New Zealand National Electronics Conference).*, Pp. 55–60.

LIST OF ABBREVIATIONS

AN	Alpha-Numeric
BB	Branch and Bound search technique
BEAM	Beam search technique
CE	Constrained Endpoint DTW
CEP	Cepstral
CT	Casual Training
DCEP	Dynamic Cepstral
DSP	Digital Signal Processor
DTW	Dynamic Time Warping
FCEP	First Order Dynamic Cepstral
FIR	Finite Impulse Response
FFT	Fast Fourier Transform
HMM	Hidden Markov Model
LR	Likelihood Ratio
LLR	Log Likelihood Ratio
LPC	Linear Predictive Coding or Linear Prediction Coefficients
NZ	New Zealand
OP	One-Pass DTW
PLP	Perceptual Linear Predictors
RMS	Root Mean Square
RPS	Root Power Sum
SC	Statistical Clustering
SNR	Signal-to-Noise Ratio
TL	Two-level DTW
UE	Unconstrained endpoint DTW
UEB	Unconstrained endpoint Band DTW
USA	United States of America
ZX	Zero Crossing rate
ZX/s	Zero Crossings per second
ZXR	Probability density of zero crossings

GLOSSARY

- **automatic speech recognition** - the recognition of words or sentences by computer.
- **autocorrelation** - the correlation of a signal with itself.
- **alphanumeric vocabulary** - a vocabulary used by speech recognition machines which consists of the alphabet and the numeral (0-9).
- **Bark frequencies** - a nonlinear frequency scale designed to model the frequency perception of the human ear. The Bark scale is based on the auditory critical bandwidth and is one representation of the frequency scale of the basilar membrane.
- **Bernoulli effect** - the suction effect caused by a sudden drop of pressure. This effect occurs when air passes through the vocal cords. Changes in air pressure at the vocal cords causes the cords to open and shut producing sounds.
- **bigram** - a measure of the likelihood of two particular words occurring consecutively.
- **consonants** - a speech unit classification.
- **connected recognition** - recognition of speech word by word. The words usually require pauses between them.
- **continuous recognition** - recognition of speech, either sentences or phrases, by recognising the words or their meaning.
- **clustering** - a technique in which multiple representations of the same information are grouped together according to some selected parameter of similarity.
- **critical band** - a frequency band representation of how the human ear masks speech in bands.
- **dipthong** - a phoneme classification consisting of two vowel sounds such as /eI/, /ou/ etc.
- **dynamic time warping** - a method of time normalisation which uses dynamic programming.
- **discrete recognition** - the computer based recognition of individual words. For discrete recognition the words must be spoken with enough delay so that each word can be recognized individually.
- **Durbin and Levinson LPC method** - a method of calculating linear prediction coefficients via autocorrelation coefficients. This method is quicker than some because it uses the properties of the Toeplitz matrix.

- **fricatives** - a type of speech sound which is produced by noise source. Fricatives can be voiced or unvoiced.
- **formant** - this is a resonance frequency of the vocal tract. Formants for voiced sounds are produced by impulsively exciting the vocal tract.
- **glottis** - the area between the vocal cords.
- **grammar** - the correct formulation of sentences.
- **hidden Markov modelling** - a method of pattern classification which uses probabilities.
- **inter-word distance** - the distance between one word and multiple representations of different word.
- **intra-word distance** - the distance between multiple representations of the same word.
- **larynx** - the structure consisting of the voice excitation apparatus.
- **LPC (linear prediction coefficients or linear prediction coding)** - either the coefficients which model the vocal tract as a system represented by an all-pole filter or the method of calculating these coefficients.
- **linguistics** - the study of languages.
- **lexical/lexicon** - the vocabulary.
- **linear time normalisation** - a method of linearly affecting the timing of one word with respect to another word.
- **liftering** - a method of windowing in the cepstral domain.
- **morphology** - the study of the form and structure of words in a language or the study of linguistic forms and structures.
- **nasal resonance** - resonances produced via the nasal cavity
- **neural networks** - a pattern classification technique which "learns" by separating the feature space into non-linear regions based on the training applied.
- **oral resonance** - resonances produced via the mouth.
- **pattern matching** - methods which compare one pattern with another. These methods often produce a distance or similarity measure giving how close one pattern is to another.
- **pharynx** - the part of the vocal tract between the mouth and the oesophagus.
- **phonation** - the production of voiced sounds by a speaker.
- **pitch** - the rate at which the glottis opens and closes, also known as the fundamental frequency (f_0).
- **phoneme** - one of a set of sound classes used to distinguish one sound from another. The set of phonemes is divided into three categories vowels, dipthongs, and consonants.

- **phonetics** - the area of study interested in the production, perception, and analysis of speech.
- **poles** - a representation of frequency resonances in the Z-domain.
- **quefrency** - the measurement used in the cepstral domain.
- **quasi-periodic** - almost occurring with a define period.
- **supra-glottal respiratory** - the vocal area above the vocal cords including the pharynx, oral cavity and nasal cavity.
- **sublaryngeal respiratory** - the vocal area below the larynx consisting of the lungs.
- **source-filter model** - a simplified model of voice production consisting of the excitation represented as a sound source and a vocal tract represented as a filter.
- **syntactics** - deals with the grammatical arrangement of words and morphemes in sentences.
- **spectrum** - the distribution of the frequencies which make up a sound.
- **speaker-independent** - pertaining to a recognition system which recognizes words from any speaker
- **speaker-dependent** - pertaining to a recognition system which recognizes words from only one speaker, usually the speaker which has trained the system.
- **spectral tilt** - tilting of the spectrum usually caused by some form of pre- or de- emphasis. Tilting of the spectrum can be caused by artifacts of the recording apparatus or by human influences such as the speakers glottal excitation.
- **stops** - sounds produced from a complete closure along the vocal tract.
- **trachea** - the tube that conveys air from the larynx to the lungs.
- **voiced sounds** - sounds that are produced with the vocal cords vibrating.
- **unvoiced sounds** - sounds that are produced without the vocal cords vibrating. Alternative forms of excitation can be caused by either a sudden release of pressure or the forcing of air through a constriction.
- **zeros** - a representation of frequency nulls in the Z-domain.

Contents

Chapter 1

INTRODUCTION

Digital Signal Processors (DSPs) have revolutionised many engineering applications. Signal processing techniques once considered too computationally intensive for real-time applications are now possible due to DSP technology. Their special purpose nature in the field of signal processing has given increased computational power to many fields of research including image processing and restoration, signal communications, signal recovery, biomedical applications and speech processing.

In the field of speech processing DSP technologies have allowed the practical real-time implementation of extremely complex, computationally intensive schemes. This is particularly so for word/speech recognition algorithms. For a word/speech recognition algorithm using a database as small as 100 words each of 0.5 seconds duration, up to 2.5 Million operations per recognition may be required for the classification algorithm alone. Many speech algorithms are still too computationally intensive to operate in real-time on today's DSPs; these algorithms include multi-speaker real-time word recognition and continuous speech recognition. However it is possible to design, with some constraints, a fast, accurate, speaker-dependent word recognition system. In order to constrain a recognition system to operate in real-time on a DSP each stage of the recognition process must be optimised, producing the best performance at the lowest computational cost. This thesis discusses four areas of a recognition system that affect both recognition speed and accuracy; pre-processing, feature extraction, pattern matching and classification, and distance measure algorithms are each investigated. A recognition scheme incorporating the optimum conditions for each stage is constructed and tested. Results are reported for both New Zealand and American accented speech.

The chapters of this thesis are arranged as follows. Chapter 2 gives an overview of speech sounds. The phonetic and linguistic characteristics of speech are described. This chapter also describes the physiology of the production mechanisms of speech. The reader is introduced to some of the terminology used later in the thesis.

Chapter 3 gives a concise history of speech analysis as it relates to speech recognition. The vast quantity of research in this area is reviewed showing the disorganised and random manner that much of this research has followed. The first recognition system was implemented as far back as 1950 giving some idea of the length of time over which speech and word recognition has been studied. However, even with this 40 year history of research, recognition systems still show limited success. A few 'cries in the wilderness' have been made to try to get a clearer understanding of the usefulness of techniques and capabilities of recognition schemes. But, as can been seen from this chapter, standardisation of techniques and comprehensive comparisons between techniques have been slow to emerge. Comparison of the various techniques used for recognition has only begun in the last 5 years, made possible by the standardisation of recognition procedures and more particularly with the standardisation of speech databases such as those produced by the National Institute of Standards and Technology (NIST), obtainable since the mid 1980s. It is in this context of comprehensive

comparison, and based on exhaustive fundamental experiments, that this thesis makes its major contribution to the field of research in speech recognition.

Chapters 4, 5 and 6 examine many of the major techniques used for recognition. Chapter 4 is a review of the features that are often proposed for word recognition, discussing the selection of features to be used in the experiments reported in Chapter 8. The features selected represent both the time and frequency information of the speech encompassing linear prediction coefficients, perceptual linear predictors, cepstral coefficients, transitional cepstral coefficients, zero-crossing analysis and energy. This set of features is chosen because they can be easily programmed to operate in real-time and because, although they are often used for speech recognition systems, they are rarely systematically compared. Chapter 5 discusses the two major time alignment methods used in recognition, dynamic time warping(DTW) and hidden Markov modelling(HMM). These two methods are examined and shown to be very similar. This chapter concludes with a discussion of why DTW was chosen over HMM for the recognition system. Chapter 6 reviews various distance measures used for speech/word recognition, examining the problems associated with choosing a distance measure which complements the choice of feature. An examination and comparison of the Mahalanobis, root power sum, and unity weighted Euclidean distance measure is undertaken. For each of chapters 4, 5 and 6, where appropriate, a discussion is also given on the effects of real-time processing to explain any deviations from the usual implementation. These deviations may affect which methods are chosen, how the method is implemented or may introduce special requirements necessary to speed up the operation of the algorithms.

The design of the complete word recognition algorithm for real-time implementation is discussed in Chapter 7. The reasons why particular methods were chosen and algorithms used are presented.

Chapter 8 provides results of recognition tests with both speaker dependent and speaker independent trials showing the effect that different pre-processing techniques, features, distance measures, speakers, accents, and reference templates have on recognition accuracy. An analysis of different pre-processing techniques shows those methods which significantly affect the accuracy and speed of the word recognition scheme, and gives those methods which produce optimal performance. This chapter constitutes the major original contribution of this thesis.

Chapter 9 examines two continuous recognition systems. Both these systems use DTW isolated word recognition techniques. A comparison of the two methods of continuous recognition using DTW are presented. First a phoneme segmentation method is examined. Problems inherent in this method such as finding beginning and ending of phonemes are discussed. The second method uses a *one-pass* DTW algorithm bypassing the need to automatically segment words. This method discusses problems associated with the variation of phoneme lengths.

Chapter 10 concludes and discusses the work presented in this thesis. Suggestions for future work are also presented in Chapter 10. Four areas of further work are discussed; the further examination of distance measures to optimise recognition accuracy, the effects of database recording parameters on the recognition accuracy, the investigation of the effects of accent on features extracted and recognition accuracy, and the continuation of continuous word recognition using DTW methods.

Chapter 2

MODELLING SPEECH

In order to model speech effectively it is important to correctly understand the anatomical and acoustical features of the vocal apparatus and its operation. This chapter firstly discusses features important to the modelling of the vocal tract and vocal cords. In the second section actual models of the vocal apparatus are examined. The models examined include acoustic, electric and digital representations. References from which the information about the modelling of the vocal tract and cords has been gathered are Fant(1971), Markel and Grey(1976), and Rabiner and Schafer(1978).

2.1 BASIC PRINCIPLES

2.1.1 Anatomy

The respiratory system can be divided into three main parts - the larynx, the supraglottal respiratory system (upper vocal area), and the sublaryngeal respiratory system (lungs and trachea). These parts are shown in Fig. 2.1

The larynx (see Fig. 2.2) is the source of quasi-periodic energy that excites the

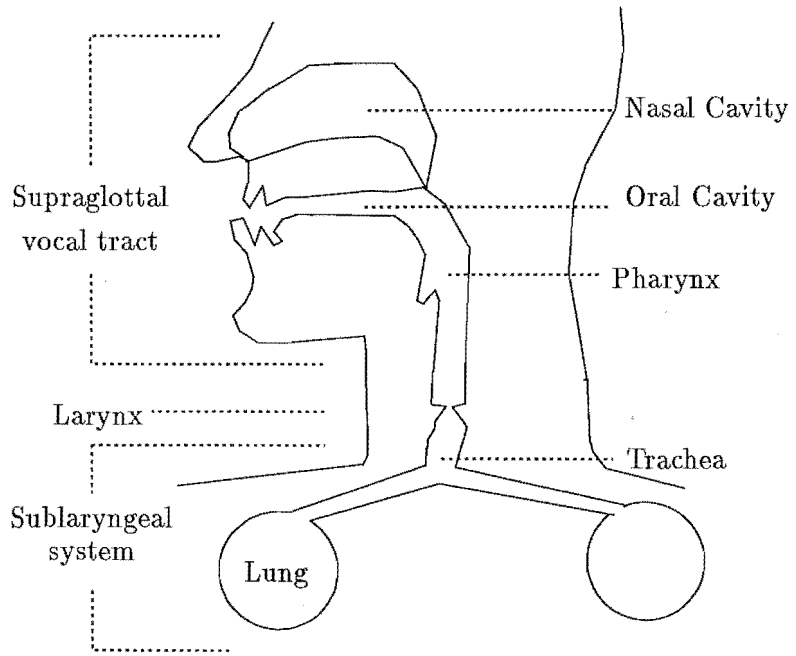


Figure 2.1. Illustration of the vocal apparatus showing the larynx, the supraglottal respiratory system and the sublaryngeal respiratory system.

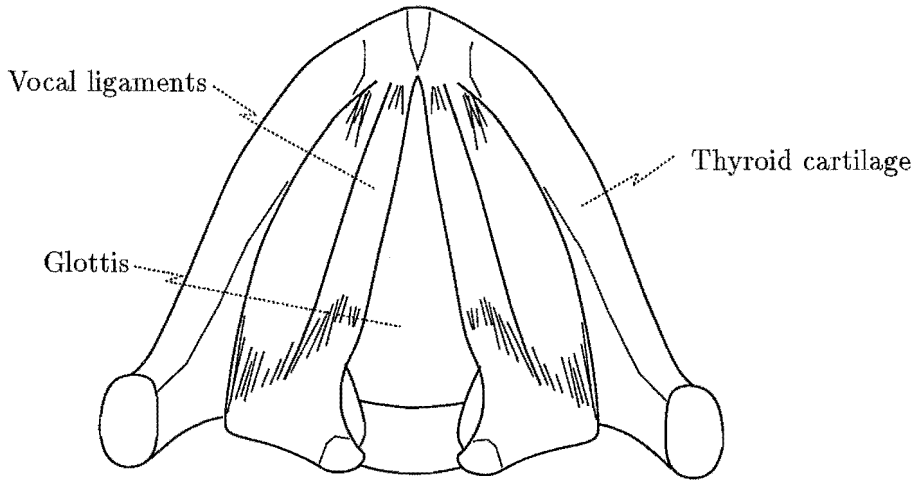


Figure 2.2. Diagrammatic view of the larynx (Gobl,1989)

vocal tract in the production of voiced sounds ((§2.1.2)). It is located at the top of the trachea, opening in to the pharynx (at the back part of the oral cavity). Externally very little laryngeal structure can be seen except for a prominence in the middle of the male neck called the *pomum Adami* (adams apple). The larynx is a complex structure of cartilages linked by muscles, ligaments and membranes. The *vocal cords*, each consisting of muscle tissue and a ligament, are set within the larynx. The area between the vocal cords is known as the *glottal opening*. The anterior position of the glottis (that parts towards the front of the neck) is the portion usually responsible for phonation. A set of muscles interconnecting the cartilages which lie around the vocal cords act to adduct or abduct the vocal cords and so enlarge or decrease the size of the glottal opening. The thickness or cross section of the vocal cords is also a function of the tension applied to them. This is important because the frequency at which the vocal cords vibrate during voicing is a function of both the tension of the vocal cords and the subglottal pressure.

Evolutionally the larynx is much older than the parts of the brain and muscle structure responsible for articulation and language. The major function of the larynx is to protect the lungs from ingestion of solids or liquids during feeding and to admit only air. To do this the larynx is able to massively adduct, sealing air in the lungs and stopping solids from entering.

The supraglottal respiratory system consists of the oral pharynx, nasal pharynx, nose, and mouth. The jaws, lips, and tongue can all be used to modify the interior shape and volume of the mouth, while the velum can open or close the nasal cavity. It is usual to designate the supraglottal respiratory system and the larynx as the *vocal tract*. The phonetic quality of differing vowel and consonant sounds is primarily due to the configuration assumed by the vocal tract.

The sublaryngeal respiratory system consists of the lungs and the trachea. The lungs serve mainly to force air through the trachea to the larynx and although they must have some acoustic effect on the sounds produced, their acoustic properties will not be discussed further in this thesis.

2.1.2 Acoustics

Sound is produced when a steady flow of air from the lungs is segmented at the larynx level by the vocal cords into a series of air puffs. The air puffs, which occur at the

fundamental frequency or pitch, cause harmonics in cavities of the vocal tract.

The vibrations of the larynx are controlled by the various laryngeal muscles, together with the pressure produced by the lungs. The larynx first acts to adduct the vocal cords across the airway. As the sublaryngeal pressure increases the resistance of the adducted vocal cords is overcome and they open momentarily, releasing a puff of air. The rapid increase of air flow when the cords open results in a drop of pressure and a consequent suction effect, known as the Bernoulli effect, which draws the cords back into a closed position. This pressure differential, plus the static tissue force of the ligaments and muscles surrounding the cords, force the folds to shut. The cycle repeats to produce a periodic flow of air causing a pressure wave known as the glottal waveform. This glottal waveform provides the acoustic excitation force for the vocal tract.

The frequency of the glottal waveform is known as the *pitch.* Its value depends upon a balance between the pressure below the cords and the static tissue force of the ligaments and muscles surrounding the cords. This balance is largely independent of the upper airway but can be controlled by the speaker who is able to alter pitch at will (Witten, 1982; Moore, 1971). The pitch of a male voice varies from as low as 50Hz up to 250 Hz with a typical value around 100Hz, while female speech ranges from 120Hz to as high as 500Hz, but typically around 200Hz.

The waveform generated by the movement of the vocal cords is not simply sinusoidal but is more closely approximated by a triangular pulse as illustrated in Fig. 2.3. This pulse train is rich with harmonics which decay at around 12db/octave. Fig. 2.3 also illustrates how the vocal cords may not completely close between cycles allowing a small air flow even in their closed position.

Glottally excited speech, known as *voiced speech*, is not the only form of vocal excitation to produce speech sounds. A second source of vocal excitation, equivalent to acoustic noise, is produced by a turbulent flow of air created at some point of stricture in the tract, and the sounds produced are known as *unvoiced speech.* The noise source for consonants may be sited at many different places in the tract – at the lips for /f/, between the tongue and teeth for /θ/, at the teeth for /s/,/ʃ/ or further back in the throat for /h/, /g/. These sounds have no pitch and hence a spectral envelope representation is broad and uniform (as shown in Fig. 2.4). The vocal cavity forward of the constriction is usually the most influential in spectrally shaping the sound. This filtering effect depends not only on the length of the tract but also on the cross sectional area of the constriction and the rate of airflow.

A third source of excitation is created by a pressure build up at some point of closure other than the vocal cords. An abrupt release of the pressure provides a transient excitation of the vocal tract. This type of excitation is used to produce sounds such as [p,b,k,t,d]. These sounds are called *plosives* and together with the sounds described in the previous paragraph, are known as *unvoiced fricatives.* Fricatives can be either voiced or unvoiced depending on whether glottal excitation is present.

The excitation signal, produced by the larynx or otherwise, is coupled to a resonant (supraglottal) system, shown diagrammatically in Fig. 2.5 as a tube opening into the vocal tract and nasal cavity. At its simplest the vocal tract can be imagined as a cylindrical pipe. For voiced speech a sound source, (the larynx), is at one end of the pipe and the other end is open representing the lips. Resonances occur in the tube with wavelengths 4L, 4L/3, 4L/5..., where L is the length of the tube (Rabiner and Schafer, 1978). For the average male the length between the vocal cords and the lips is approximately 17cm and so the resonant frequencies would occur at 500Hz, 1500Hz, 2500Hz. When excited by the harmonic rich waveform of the larynx, the vocal tract resonances produce peaks, known as formants, in the spectral envelope of the speech wave, as shown in Fig. 2.6 . The lowest formant, (F_1), usually varies between 200 to

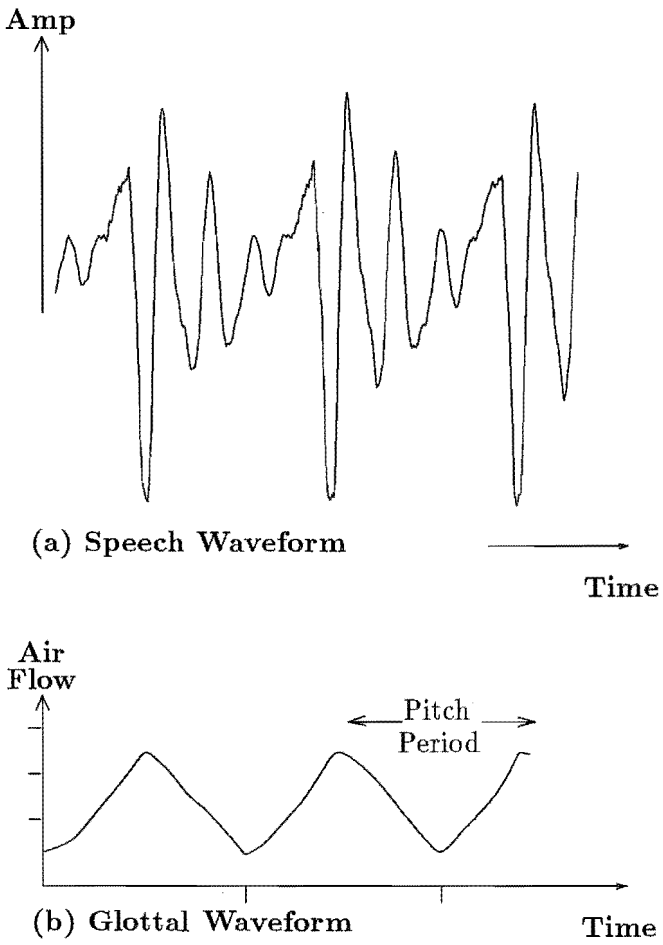


Figure 2.3. Illustration of (a) speech produced by the vocal tract excited by a (b) glottal wave (Liberman and Blumstein, 1987). The repetitive nature of the speech waveform is caused by the opening and closing of the glottis producing the glottal waveform. The time between successive openings of the glottis is known as the pitch period of the speech.

1000 Hz, formant two, (F_2), varies around 500 to 2500 Hz and formant 3, (F_3), varies around 1500 to 3500 Hz.

The cylindrical pipe model of the vocal tract represents the vocal tract as a straight segment, whereas actually it is bent into two sections with a vertical section at the rear of the mouth and a roughly horizontal section in the front. The tongue placement within these roughly right-angled sections can induce some of the largest effects on the output resonances by altering the relative length of each tube. The formant structure can also be modified by the shape and extent of the lip opening. Therefore any formant pattern of a particular sound is the outcome of the acoustic character of the whole tract working as one resonant system. Fig. 2.7 shows the different formant patterns for the three different vowel sounds [oo] as in hood, [a] as in had, [ar] as in hard showing the differences of the patterns that can be obtained with variations of lips and tongue placement.

With real speech the vocal tract is continuously moving, altering its shape, and so changing the position of resonances which change the sounds produced. A three dimensional plot of the variation of speech energy versus time (Fig. 2.8) illustrates how the formants change throughout the duration of spoken words, moving smoothly from

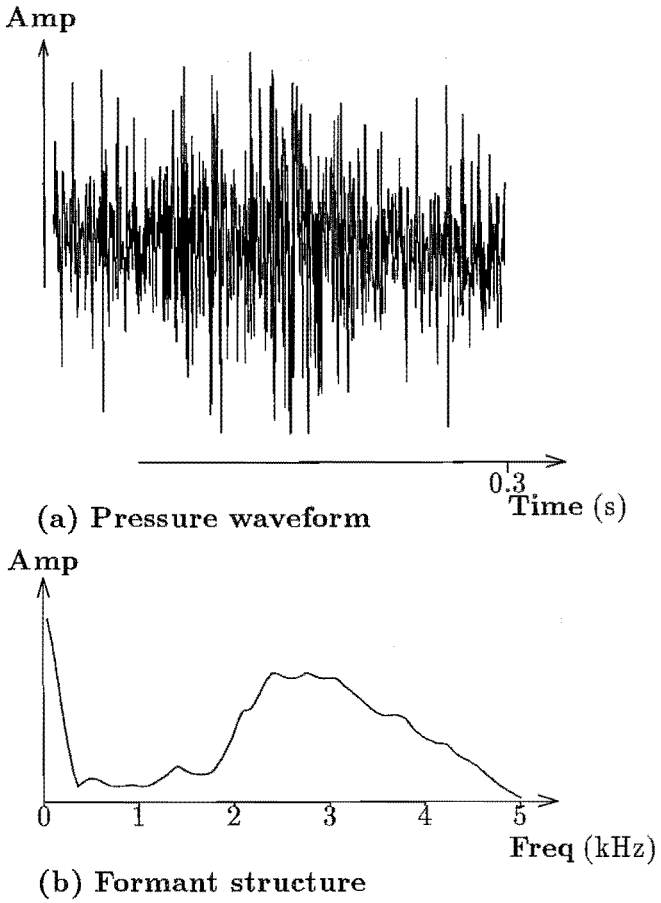


Figure 2.4. Pressure waveform and spectral envelope of unvoiced sounds showing wide band frequency information. Note that the plot shows no pitch information.

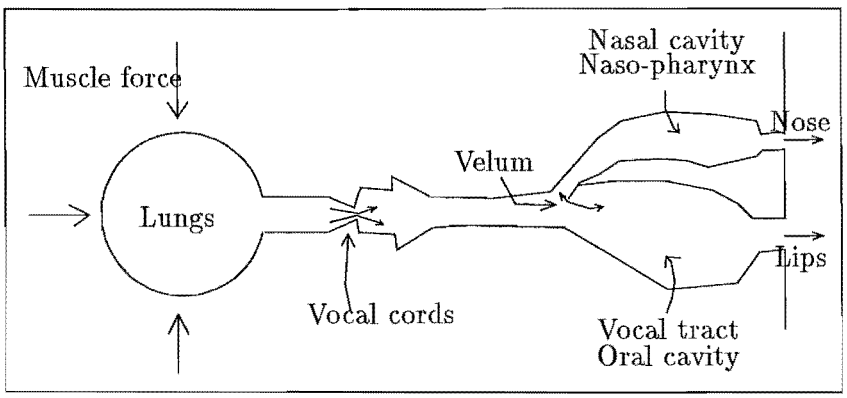


Figure 2.5. Model of the vocal tract. A sound source is caused by the lungs forcing air through the vocal cords and producing an excitation which is shaped by the vocal tract and nasal cavity. Speech is output as a sound from the lips and nose.

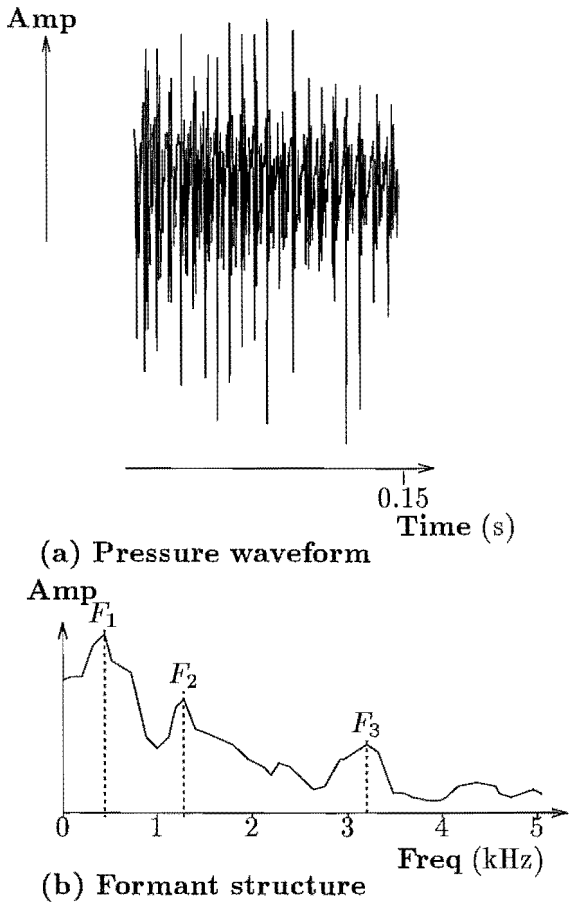


Figure 2.6. Pressure waveform and spectral envelope of voiced speech showing three formant peaks, F1, F2 and F3. Note the repetitive nature of the pressure waveform caused by the air puffs released by the glottis as it opens and closes.

one configuration to the next.

2.1.3 Phonetics

Phonetics is a term given to the field of study concerned with the description of speech in terms of basic units, each unit representing a fundamental sound called a *phoneme*. The set of phonemes can be divided into three main categories: vowels, dipthongs and consonants. Division into these categories is difficult to justify by simply looking at the way these sounds are produced. Many consonants are nearly indistinguishable from vowels in their formation and in their acoustical properties. One definition of a vowel is 'a speech sound which may constitute a syllable or the nucleus of a syllable' whereas a consonant is considered 'a speech sound which is used marginally with a vowel or dipthong to constitute a syllable' (Ladefogad, 1973).

In the following sections the division of speech sounds between vowels, consonants and dipthongs is examined more closely.

2.1.3.1 Vowels and Dipthongs

Vowels consist of sounds produced solely by vocal cord excitation (voiced sounds). Normal articulation of a single vowel has a tract shape which is relatively stable and

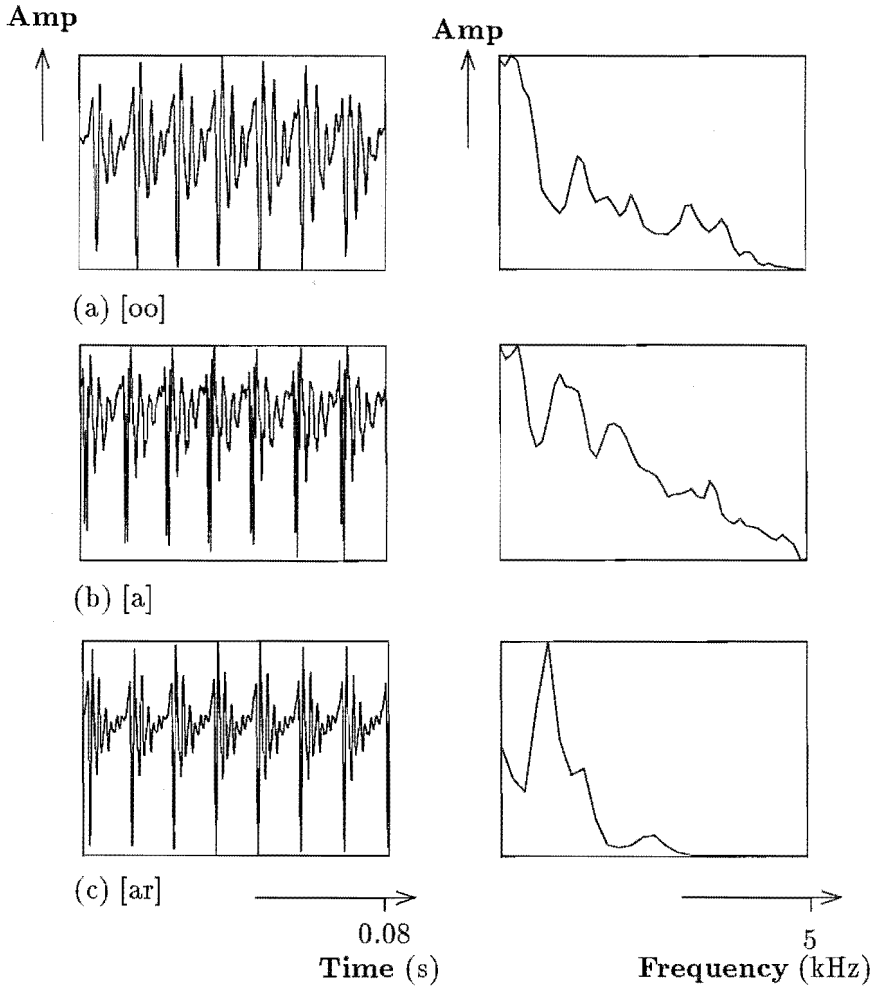


Figure 2.7. Signal and formant structure of three vowel sounds, (a) [oo] from hood, (b) [a] from had and (c) [ar] from hard. Note changes in the signal cause changes in the frequency structure and location of frequency peaks.

relatively open. Vowel sounds also have very little nasal tract coupling being produced mainly from the mouth.

The different vowel sounds are produced by changing the shape of the oral cavity. This usually occurs by moving the tongue and, to a lesser extent, by changing jaw opening and lip rounding. As the tongue movement is the most influential, the English vowels are usually categorised according to tongue position as shown in Table 2.1 (Skinner and Shelton, 1978; Kaplan, 1971).

Dipthongs are a special kind of vowel, a single phoneme consisting of two different vowel positions. One of the positions is the dominant known as the *nucleus*, it has a longer duration and greater stress. The second position is called the *glide*. These two positions are formed by movements of the vocal tract as the sound is made. The English dipthongs are */eI/* as in *made*, */ou/* as in *own*, */aI/* as in *ice*, */au/* as in *out*, and */ɔI/* as in *oil* .

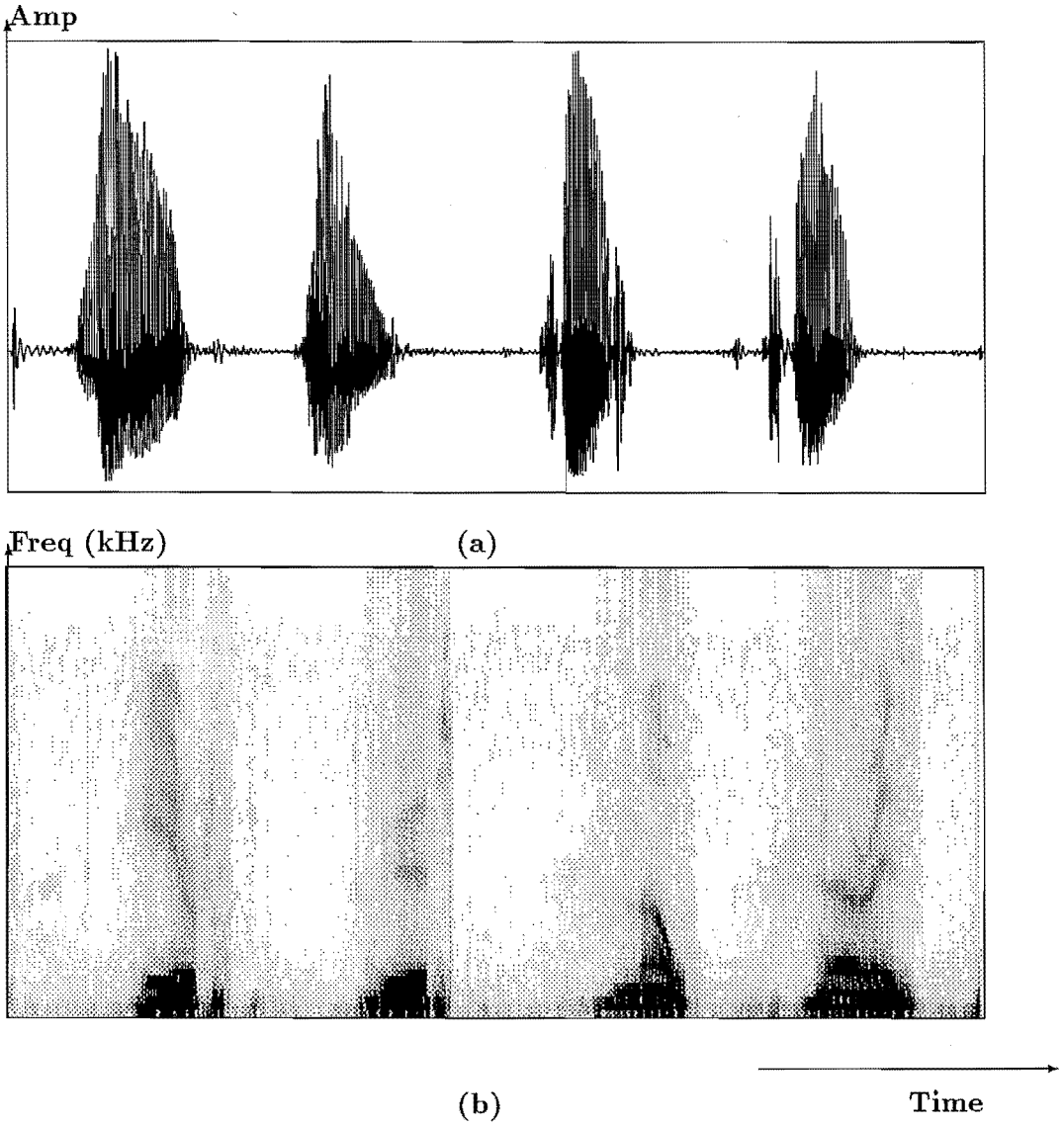


Figure 2.8. Temporal, (a), and spectrographic, (b), plots of the sentence 'ZERO ONE TWO THREE'. Time is plotted horizontally. For the spectrographic plot the frequency is plotted vertically and the energy of the sounds is represented by the darkness of the plot. As the time waveform slowly changes shape the formant structure of the spectrogram also changes shape with time.

2.1.3.2 Consonants

Consonants make up those sounds that are neither vowels or dipthongs. They can be formed by either voiced or unvoiced excitation or as a nasalised sound. Consonants are usually categorised by the part of the speech mechanism most important in the production of the sound or type of excitation used. Using this criterion, consonants can be represented by five major classes - stops, fricatives, affricatives, oral resonants, and nasal resonants. These consonants are outlined in Table 2.2 (Levelt, 1989)

2.1.3.3 Other Sub-Word Units

Phonemes, although the smallest sub-word unit, are not the only word reduction unit. Diphones, syllables, demi-syllables and disyllables are also sub-word units, breaking

DEGREE OF CONSTRICTION	TONGUE HUMP POSITION		
	FRONT	CENTRAL	BACK
HIGH	/i/ *eat*	/ɝ/ moth*er*	/u/ s*uit*
	/I/ *it*		/U/ b*oo*k
MEDIUM	/e/ v*a*cation	/ə/ sof*a*	/o/ *o*bey
	/ɛ/ *e*ver	/ʌ/ *u*p	/ɔ/ l*aw*
LOW	/æ/ *a*t		/ɒ/ n*o*t
	/a/ cl*a*ss		/ɑ/ f*a*ther

Table 2.1. Vowels of the English language (Skinner and Shelton,1978).

Graphemic :	èmigrante
phonemes :	e m i g R ã t
diphones :	e em mi ig gR Rã ãt t
syllables :	e mi gRãt
demi-syllables :	e em mi ig gRã ãt t
disyllables :	e emi igRã ãt

Figure 2.9. Representation of the word *emigrate* by sub-word units (Mariani,1989).

the speech into primary units. An example of how these sub-word units represent a word is shown in Fig. 2.9. Detailed explanation of each subword unit may be found in Mariani(1989).

2.2 ACCENTS

A speaker organises his/her utterances in patterns of stressed and unstressed syllables, assigning various degrees of stress (or accent) to different syllables. In such cases various accents produce various patterns of stressed and unstressed sounds. This rhythm of speech is reflected in part by the duration of the syllables, where stressed syllables are longer. Stress is also realised by variations in amplitude (or intensity) and pitch (refer (§2.1.2)) (Levelt, 1989). Speech quality is also important to speech accent. Often qualities such as 'nice' or 'musical' or 'ugly' or 'harsh' are used to describe a speaker's speech (Crystal, 1980). These qualities are due to the variations of formants during phonation, and in particularly are realised on the first two formants of the speech. Fig. 2.10 shows the difference in the major vowels of two speech accents, New Zealand and American English. The movement of the formants of the vowel sounds can produce the most difference between two speakers voicing the same speech.

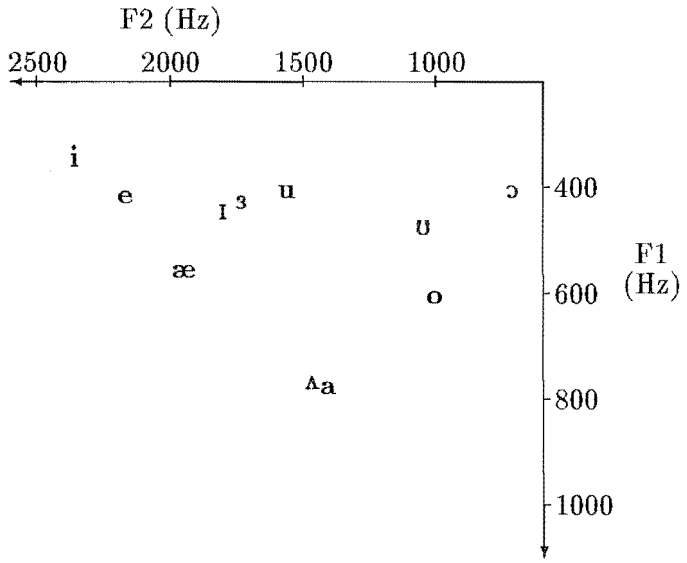
CONSONANT	DEFINITION	EXAMPLES	
		VOICED	UNVOICED
Stop or Plosive	Produced when a complete closure is made along the vocal tract. Pressure in the lungs is suddenly released producing sound.	/**b**/ be /**d**/ day /**g**/ go	/**p**/ pay /**t**/ to /**k**/ key
Fricative	Produced by an audible incoherent noise source by forcing air through a constriction along the vocal tract.	/**v**/ vote /ð/then /**z**/ zoo /ʒ/ azure	/**f**/ for /θ/ thin /**s**/ see /ʃ/ she
Affricative	Combination of a stop phase immediately followed by a fricative. The stop and fricative portion of the sound are produced at the same place of articulation.	/ʤ/ jam	/tʃ/chair
Oral Resonances	Produced by oral cavity resonances set up by the vibrational action of the vocal cords. As vowel sounds are produced in the same way these consonants are sometimes called 'semi-vowels' or 'vowelised' consonants.	/**w**/ we /**j**/ yes /**l**/ let /**r**/ red	
Nasal Resonances	These consonants are produced as oral resonants but, when producing nasals, the velum is open, permitting resonances in the nasal cavity. Also the oral cavity is completely closed off at some point thus forcing air through the nasal opening.	/**m**/ me /**n**/ new /ŋ/ sing	

Table 2.2. The English consonants (Kaplan, 1971). Represented by five classes - stops, fricatives, affricatives, oral resonances, and nasals.

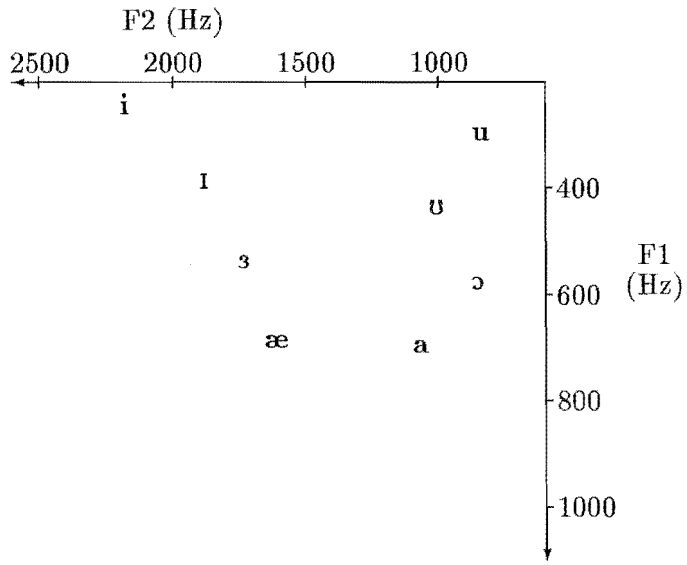
2.3 MODELLING THE VOCAL TRACT

In this section the vocal tract and speech producing organs are represented as an acoustical system. A model of the vocal tract is examined which attempts to model the physical processes which affect the acoustics of the output speech. In order to represent these processes a complicated description of the vocal tract model is needed.

The second part of this section examines another representation, requiring much less information, which models the vocal tract at a more abstract level. This abstract description is the *source-filter model* forming the basis of an electrical analog of the vocal tract, first investigated by Fant (1960).



(a) New Zealand English



(b) American English

Figure 2.10. The first (F1) versus the second (F2) formant positions of various vowel sounds for two accents, (a) New Zealand (Maclagan,1982) and (b) North American English (Ladefogad,1982). Formant positions are shown for a typical male speakers.

2.3.1 Direct Modelling of the Vocal Tract

To accurately model the vocal tract as an acoustical system, the factors which influence the output must be examined. These factors are listed by Rabiner and Shafer (p.57) as;

- TIME VARIATIONS OF THE VOCAL TRACT. For the vocal tract to be able to vary its resonant frequencies, and hence the sounds produced, it must vary its configuration in time. For a single vowel sound, however, the movement of the

vocal tract is minimal and is usually neglected when modelling such a sound.

- LOSSES DUE TO HEAT CONDUCTION AND VISCOUS FRICTION AT THE VOCAL TRACT WALLS. Energy is lost by the air in the tract due to both heat conduction through the walls of the vocal tract and through viscous friction between the air and the tract walls.

- SOFTNESS OF THE VOCAL TRACT. The walls of the vocal tract vibrate due to the varying forces along the tract, causing changes or perturbations in the tract which affect the resonances of the tract.

- RADIATION OF SOUND AT THE LIPS. Termination of the vocal tract at the opening of the lips or nostrils places a radiation load at the output. At low frequencies the effect is negligible (that is the radiation impedance approximates a short circuit termination), whereas at high frequencies the effect is approximately equal to the radiation resistance at the lips (Rabiner and Schafer, 1978, p71). This has the effect of de-emphasising the output volume velocity at higher frequencies.

- NASAL COUPLING. For nasal consonants the oral tract is completely closed and sound radiation is from the nose, while for nasalised vowels sound is produced from both the nasal and oral passages. For nasalised vowels the speech signal is then a superposition of nasal and oral outputs. For nasal sounds the closed oral cavity can trap energy at certain frequencies, preventing those frequencies from appearing in the nasal output. (Rabiner and Shafer, 1978, p78). This results in anti-resonances as well as resonances being present in the frequency response. Another frequency response effect is that for nasalised sounds the nasal formants having broader bandwidths than the non-nasal sounds. This is attributed to the greater viscous friction and thermal loss due to the large surface area of the nasal cavity.

- EXCITATION OF SOUND IN THE VOCAL TRACT. All controlling forces for the movement of the vocal cords need to be considered for accurate representation. These forces are firstly due to pressure of air forced through the cords by the lungs, secondly due to the tension and stiffness of the cords, and finally due to the area of glottal opening.

It is computationally impractical to include all these factors into a model. Usually only those effects which most grossly affect the speech are considered. A close representation models the vocal tract as a tube of varying cross sectional area, shown in Fig. 2.11, with yielding walls but with no losses due to heat conduction and friction. These assumptions are reasonable and only produce significant errors at frequencies below 500Hz (Rabiner and Schafer, 1978). The Portnoffs equations (Rabiner and Schafer, 1978) can be used to approximate the acoustic system by using the laws of conservation of mass, momentum and energy, producing the momentum equation

$$-\frac{\partial p(x,t)}{\partial x} = \rho \frac{\partial U(x,t)/A(x,t)}{\partial t}, \tag{2.1}$$

and the continuity equation

$$-\frac{\partial U(x,t)}{\partial x} = \frac{1}{\rho c^2}\frac{\partial (p(x,t).A(x,t))}{\partial t} + \frac{\partial A(x,t)}{\partial t}. \tag{2.2}$$

where $A(x,t)$ is the cross-sectional area, that is the area normal to the longitudinal dimension, of a tube. $p(x,t)$ is the sound pressure, c is the velocity of sound, $U(x,t)$ is the volume velocity, and ρ is density of air.

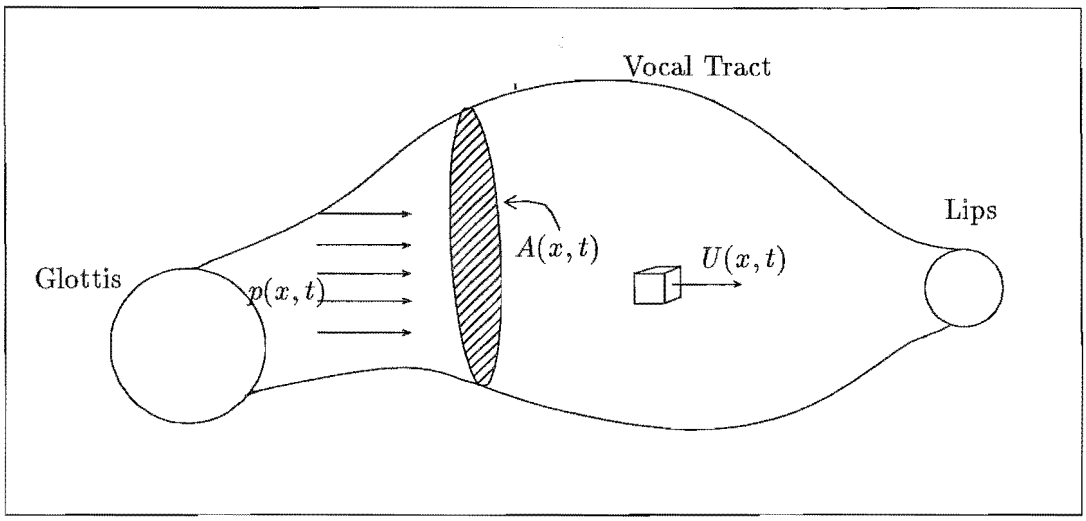


Figure 2.11. Model of the vocal tract as a tube with varying cross sectional area $A(x,t)$. Also illustrated is the volume velocity $U(x,t)$ and the pressure wave $p(x,t)$.

A solution to these equations is complicated, requiring extensive numerical analysis. One problem is that the initial tract area function $A(x,0)$, must be known. Accurate measurements of this initial area function have been tried using direct methods such as X-rays (Fant, 1960), or indirect methods, such as exciting the vocal tract with external impulses (Sondhi and Gopinath, 1971), or by measuring the acoustic properties of the speech (Mermelstein, 1967). Another solution is to remove the problem of finding the area function by modelling with the vocal tract set in a constant configuration, that is by setting $A(x,t) = A(x)$. The partial differential equations, (2.1) and (2.2) then reduce to (Markel and A.H. Gray, 1976),

$$\frac{\partial p(x,t)}{\partial x} = \frac{-\rho}{A(x)} \frac{\partial U(x,t)}{\partial t}, \tag{2.3}$$

and

$$\frac{\partial U(x,t)}{\partial x} = -\frac{A(x)}{\rho.c^2} \frac{\partial p(x,t)}{\partial t}. \tag{2.4}$$

These equations combine to give the classical Webster horn equation (Markel and A.H. Gray, 1976) which models the vocal tract as a smoothly flaring horn as follows

$$\begin{aligned} \frac{\partial^2 p(x,t)}{\partial x \partial t} &= \frac{-\rho}{A(x)} \frac{\partial^2 U(x,t)}{\partial t^2}, \\ &= -\rho c^2 \frac{\partial .}{\partial x} \left[\frac{1}{A(x)} \frac{\partial U(x,t)}{\partial x} \right]. \end{aligned} \tag{2.5}$$

such that

$$\frac{\partial}{\partial x} \left[\frac{1}{A(x)} \frac{\partial U(x,t)}{\partial x} \right] = \frac{1}{c^2 A(x)} \frac{\partial^2 U(x,t)}{\partial t^2} \tag{2.6}$$

Solving these equations for the output volume velocity for a vocal tract response for a constant position gives a frequency response like that plotted in Fig. 2.12. This response contains an infinite number of poles equally spaced on the frequency axis.

A more realistic model includes the effect of tract losses, such as non-static vocal walls, and less pronounced effects such as viscous friction and thermal conduction. The effect of yielding walls gives an output volume velocity response with damped resonances. These damped resonances have centre frequencies which are slightly higher

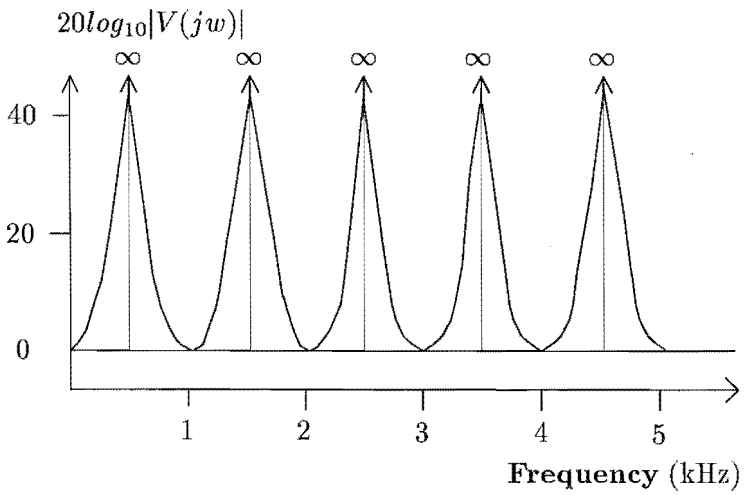


Figure 2.12. Frequency response for a flared horn with no losses. A flared horn without losses gives a basic model of the vocal tract response.

than the undamped case, while having bandwidths which are wider. With the addition of friction and thermal losses to the model the bandwidths are further increased. Adding lip impedance effectively de emphasises the volume velocities at high frequencies effectively reducing the magnitudes of the higher frequency volume velocity components(Rabiner and Schafer, 1978).

2.3.2 The Vocal Tract as a Series of Uniform Tubes

For many systems, such as word recognition systems, requiring the modelling of speech a high degree of accuracy is not required. The real problem is to model the speech signal efficiently, representing the relevant information in the sound with the least number of parameters. One efficient representation of the vocal apparatus is the source-filter model where the sound source (providing the excitation) is separated from the vocal tract filter. A diagram of this model is drawn in Fig. 2.13. An important assumption of the representation is that there is no coupling between the vocal tract and sound source.

The excitation or sound source of this model must produce two basic excitations, a periodic pulse train for voiced sounds and a noise-like excitation for unvoiced sounds. The pulse train is most simply represented as a train of impulses filtered by a glottal shaping filter. For voiced speech, the excitation source produces a waveform whose frequency components decay at about 12dB/octave. To represent this function often a sawtooth or triangular wave is employed as their shape represents that of an actual glottal pulse produced by the vocal cords. Because the sawtooth waveform exhibits discontinuities it has the wrong asymptotic rate of decay (6dB/octave instead of 12 dB/octave) (Witten, 1982) a better representation is the triangular pulse which does decay at 12 dB/octave. The second type of excitation (to produce unvoiced sounds) requires only a noise generator.

The vocal tract model approximates the vocal tract as a series of uniform tubes. An accurate representation of the vocal tract can be made if a large number of tubes are used. However, only 5-12 tubes are used in practice. Fig. 2.14 shows a diagram of a model with 6 tubes.

The following assumptions about the vocal tract are made for the uniform tube model (Markel and A.H. Gray, 1976)

- The vocal tract is represented as M connected acoustic tubes of equal length.

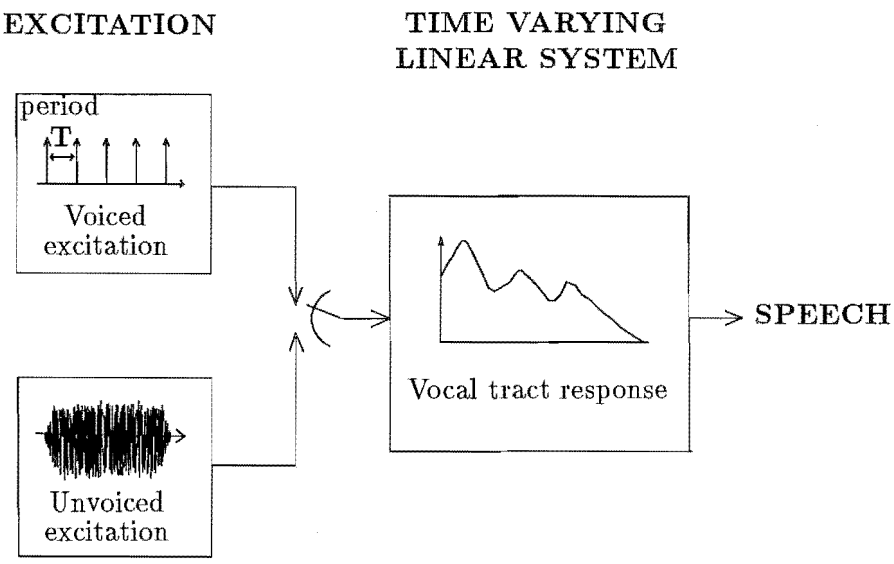


Figure 2.13. Model of the speech production system with the excitation decoupled from the vocal tract filter (the source-filter model).

Each tube has its own volume and hence a unique transfer function.

- Propagation through an individual tube section is treated as a plane wave.
- All losses due to wall vibration, viscosity and heat conduction are ignored.
- The vocal tract model is completely decoupled from the excitation source model, allowing constant boundary conditions.
- Effects of the nasal tract are ignored.

Taking all these approximations into account the wave propagating within the tube can be found by solving Portnoffs equations (2.1) and (2.2) for each section, where $U_m(x,t)$ and $P_m(x,t)$ are the volume velocity and the pressure respectively in section m. The index m is taken from the lips ($m = 0$) to the glottis, ($m = M-1$) (Wakita, 1973). At the mth tube, which has cross-sectional area A_m the pressure and volume velocity has the form (Wakita, 1973)

$$P_m(x,t) = \rho\frac{c}{A_m}[U_m^+(t - x/c) + U_m^-(t + x/c)], \tag{2.7}$$

and

$$U_m(x,t) = [U_m^+(t - x/c) - U_m^-(t + x/c)]. \tag{2.8}$$

This solution represents a one dimensional wave equation for the volume velocity and pressure expressed as a linear combination of forward (+) (from glottis to lips) and backward (-) travelling waves.

Equations (2.7) and (2.8), allow the volume velocity (or pressure) at any location, x, and time, t, within the mth section to be found as the difference between the forward travelling wave and the backward travelling wave.

The areas A_m of the cylinders can be calculated from these equations by relating pressure and volume at the boundaries. An illustration of the positive and negative going volume velocities at a boundary is shown in Fig. 2.15.

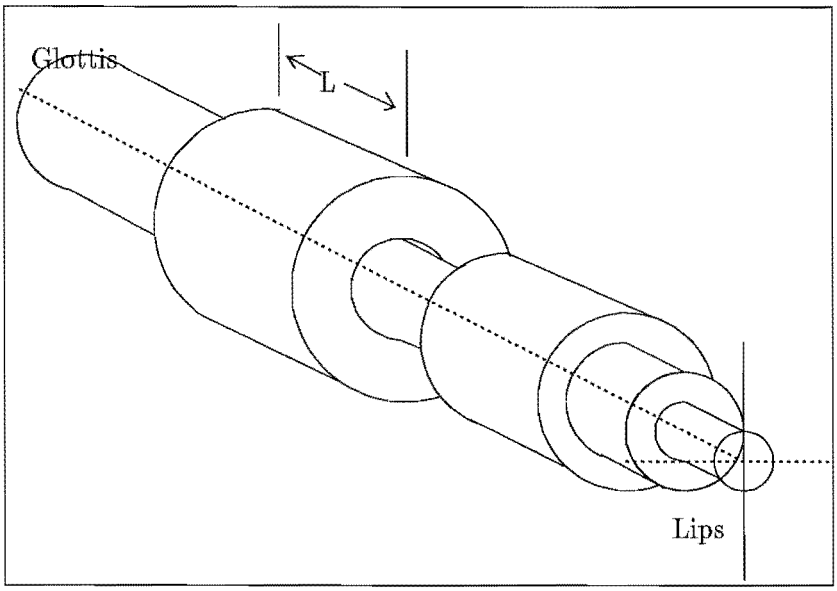


Figure 2.14. Vocal tract modelled as a series of M tubes, where in this case $M = 6$. The tubes have constant length L but different cross-sectional area. By changing the cross-sectional area of the tubes various vocal-tract shapes can be produced, representing those shapes of the vocal-tract required to produce sounds.

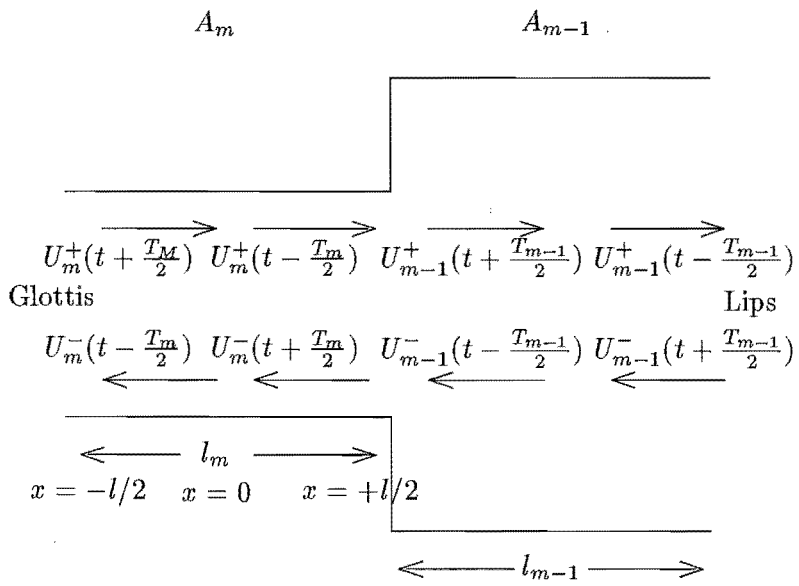


Figure 2.15. Junction between two lossless tubes m and $m-1$ of length l_m and l_{m-1}, showing positive going and negative going travelling waves. At the boundary the pressure and volume velocity must be continuous in both time and space. The time to travel through the mth tube is represented as T_m and the $m-1$ tube is T_{m-1}. Generally the length of the tubes is kept constant such that $l_m = l_{m-1} = l$ and in such cases $T_m = T_{m-1} = T$

At a boundary between mth and $(m-1)th$ section, the pressure and volume velocities must be continuous such that

$$\begin{aligned} P_m(l/2,t) &= P_{m-1}(-l/2,t), \\ U_m(l/2,t) &= U_{m-1}(-l/2,t). \end{aligned} \tag{2.9}$$

where x is measured from the centre of each section such that a section of length l spans from $-l/2$ to $+l/2$. Substitution of these equations into equations (2.7) and (2.8) gives,

$$\begin{aligned} U_m^+(t-\tau) - U_m^-(t+\tau) &= U_{m-1}^+(t+\tau) - U_{m-1}^-(t-\tau), \\ U_m^+(t-\tau) + U_m^-(t+\tau) &= [U_{m-1}^+(t+\tau) + U_{m-1}^-(t-\tau)][A_m/A_{m-1}] \end{aligned} \tag{2.10}$$

where τ is the time to travel through half a section of the tube and is equal to $T_m/2$.

At the $m-1$ section, in the absence of a reverse-travelling wave ($U_{m-1}^-(t-\tau) = 0$), $-U_m^-(t+\tau)$ can be considered the reflection of $U_m^+(t-\tau)$. Therefore combining the two equations above allows areas to be calculated for the vocal tract sections which can be generalised so that the amount of $U_m^+(t-\tau)$ that is reflected at any boundary can be calculated as:

$$-U_m^-(t+\tau) = [A_{m-1} - A_m]/[A_{m-1} + A_m]U_m^+(t-\tau), \tag{2.11}$$

and a reflection coefficient can be defined for any boundary as:

$$r_m = [A_{m-1} - A_m]/[A_{m-1} + A_m]. \tag{2.12}$$

This equation is easily rearranged so that

$$\frac{A_m}{A_{m-1}} = \frac{1 - r_m}{1 + r_m}. \tag{2.13}$$

This relationship allows the relative cross-sectional areas of the vocal tract tubes to be calculated from reflection coefficients. Note that the reflection coefficients are able to be calculated directly from the speech signal via linear prediction coefficients (as discussed more fully in §4.3). The calculated areas are relative areas and typically this is achieved by setting the area of the first or last tube to unity so that all other areas are evaluated relative to it. This normalized cross-sectional area of the model vocal tract is called the *area function*. For vocal tract shapes to be calculated from particular sounds it is worth noting that the area function is not unique. Multiple area functions can have the same transfer function (such as the area function and its inverse). Further more the area function depends critically on the formant bandwidths. That is one sound can be produced with the same formant locations but differing bandwidths. The area function derived from these sounds can differ markedly (Sondhi, 1979). A third difficulty is that the transfer function of the vocal tract cannot be estimated reliably at frequencies higher than 3kHz (Elder, 1991). Although the area functions do not have unique correspondence to the shape of the vocal tract during voicing, the shape usually generated (if pre-emphasis is used prior to linear predictive analysis) appears to be reasonable estimates of the vocal tract shape (Elder *et al.*, 1987) From a practical point of view, it is irrelevant whether the shapes are non-unique if reasonably accurate vocal tract representations are generated from actual speech signals (Elder, 1991).

To produce a simple mathematical description of the acoustic tube vocal tract model there is required further manipulation of the volume velocity equations which must be obtained in terms of the reflection coefficients. The reflection coefficients can be substituted into the volume velocity equations such that the forward and backward

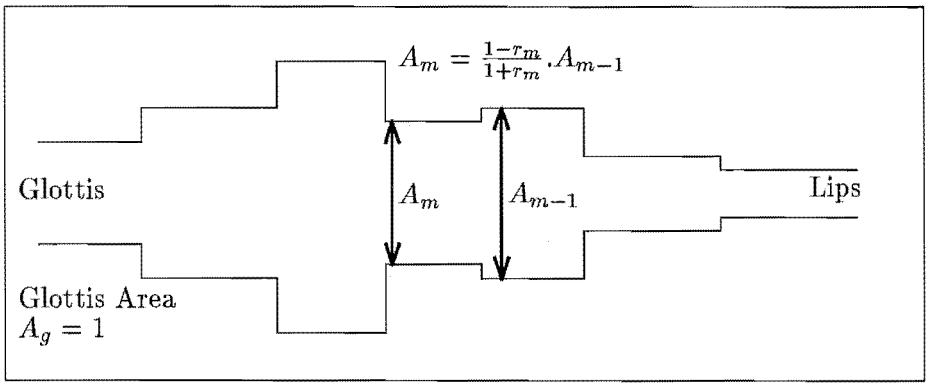


Figure 2.16. Acoustic tube representation of the vocal tract.

volume velocity at the boundary of sections m and $m-1$ can be found in terms of the reflection coefficients,

$$U_m^+(t-\tau) = \frac{U_{m-1}^+(t+\tau) - r_m U_{m-1}^-(t-\tau)}{1+r_m}, \tag{2.14}$$

and

$$U_m^-(t+\tau) = \frac{-r_m U_{m-1}^+(t+\tau) + U_{m-1}^-(t-\tau)}{1+r_m}. \tag{2.15}$$

Solving (2.14) for the forward-travelling volume velocity, $U_{m-1}^+(t+\tau)$, at the left edge of section $m-1$, and substituting this into 2.15 to obtain the reverse-travelling volume velocity wave, $U_m^-(t+\tau)$, at the right edge of section m, gives the equations,

$$U_{m-1}^+(t+\tau) = r_m U_{m-1}^-(t-\tau) + (1+r_m) U_m^+(t-\tau) \tag{2.16}$$

and

$$U_m^-(t+\tau) = (1-r_m) U_{m-1}^-(t-\tau) - r_m U_m^+(t-\tau) \tag{2.17}$$

At the glottis boundary $U_{M-1}^+(t+\tau)$ and $U_{M_1}^-(t-\tau)$ must be related, while at the lips $U_0^+(t-\tau)$ and $U_0^-(t+\tau)$ must be related. At the boundary where lip termination occurs the lips are defined as being open, analogous to a short circuit and such that the pressure at the lips is zero. The boundary condition is therefore

$$U_0^+(t-\tau) = -U_0^-(t+\tau) \tag{2.18}$$

The volume velocity output of the acoustic tube at the lips, $U_L(t)$, is

$$U_L(t) = U_0^+(t-\tau) - U_0^-(t+\tau) \tag{2.19}$$

which results in

$$U_L(t) = 2U_0^+(t-\tau) \tag{2.20}$$

Determination of the boundary condition at the glottis is more involved. A model for the glottis is that of an acoustic tube driven by a volume velocity source $U_G(t)$ whose source impedance is Z_G. Markel and Gray (1976) present two models for the boundary conditions at the glottis. One of which defines an artificial section M whose impedance is matched to the volume velocity source impedance Z_G. The model containing the artificial impedance section, defined as section M, is precisely equivalent to the model without an added section (whose section $M-1$ is directly connected to the glottal volume velocity source). However the extra section adds a delay which is compensated

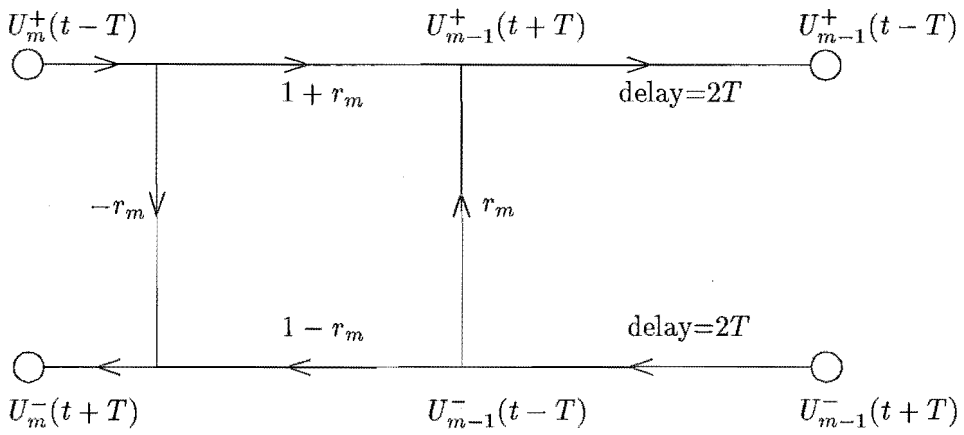


Figure 2.17. Signal flow-graph representing the junction between two lossless tubes.

for by defining the glottal source volume velocity as $U_G(t+2\tau)$. Assuming Z_G is real, Markel and Grey derive a source volume velocity $U_G(t)$ such that

$$U_G(t) = \frac{2U_{M-1}^+(t+\tau) - 2r_M U_{M-1}^-(t-\tau)}{1+r_M} \tag{2.21}$$

and is identical to (2.14) if

$$U_M^+(t-\tau) = \frac{U_G(t)}{2} \tag{2.22}$$

The volume velocity $U_M^+(t-\tau)$ represents a forward-travelling wave in the artificial section M added to the vocal tract model. The reverse-travelling wave in section M can be determined from (2.15) with $m = M$, which results

$$U_M^-(t+\tau) = (1-r_M)U_{M-1}^-(t-\tau) - r_M U_M^+(t-\tau) \tag{2.23}$$

Signal flow-graphs can be used to represent the multiplications and additions of equations (2.16) and (2.17). A signal flow-graph representing the mathematical description of the physical model of Fig. 2.16, is shown in Fig. 2.17.

2.3.3 Obtaining Filter Parameters from Area Parameters

The tube model has many properties in common with digital filters allowing a transfer function of the complete lossless tube model to be found. The transfer function represents the vocal tract and can be written as a z-domain function if the signal is sampled at a rate of not less than $\frac{1}{2T} = \frac{1}{4\tau}$ which is the time for a signal to travel twice the length of a tube section. To understand why this is the necessary value consider a vocal tract tube model consisting of the concatenation of N tube sections. If an impulse is applied at the glottal end such that $U_G(t) = \delta(t)$. Then the impulse will propagate down the tube being partially propagated and partially reflected at each boundary. The impulse response of the tube model can be represented as (Rabiner and Schafer, 1978)

$$v(t) = a_0\delta(t-2N\tau) + \sum_{k=1}^{\infty} a_k\delta(t-2N\tau-4k\tau) \tag{2.24}$$

Thus the soonest the impulse can reach the output is $2N\tau$ secs. This time is only representative of the delay between the production of the excitation at the glottis and

the sound at the lips. However successive impulses due to reflections at the junctions reach the output at multiples of 4τ seconds later (the time to propagate both ways in one section). Because the system function of a lossless tube model has many properties in common with digital filters the model can be written in the form

$$V(s) = \sum_{k=0}^{\infty} a_k e^{-s(N+2k)2\tau} = e^{-sN2\tau} \sum_{k=0}^{\infty} a_k e^{-s4\tau k} \tag{2.25}$$

where the factor $e^{-sN2\tau}$ corresponds to the delay time required to propagate through all N sections and the quantity

$$\hat{V}(s) = \sum_{k=0}^{\infty} a_k e^{-sk4\tau} \tag{2.26}$$

is the system function of a linear system whose impulse response is $\hat{v} = v(t + 2N\tau)$. Equation (2.26) represents the resonance properties of the system. Examining the frequency response, $\hat{V}(\omega)$

$$\hat{V}(\omega) = \sum_{k=0}^{\infty} a_k e^{-j\omega k4\tau} \tag{2.27}$$

and

$$\hat{V}(\omega + \frac{2\pi}{4\tau}) = \hat{V}(\omega) \tag{2.28}$$

which is reminiscent of the frequency response of a discrete-time system, such that, if the input to the system (the excitation) is bandlimited to frequencies below $\frac{\pi}{4\tau}$ then the input can be sampled with a period 4τ and filtered with a digital filter whose impulse response is

$$\begin{aligned} \hat{v}(n) &= a_n \quad n \geq 0 \\ &= 0 \quad n < 0. \end{aligned} \tag{2.29}$$

The $z-$domain representation of the vocal tract is $\hat{V}(\omega)$ with $e^{-j\omega 4k\tau}$ replaced by z^{-1} such that

$$\hat{V}(z) = \sum_{k=0}^{\infty} a_k z^{-k}. \tag{2.30}$$

To derive a general model for a lossless tube discrete-time model for speech production the ratio of the volume velocity at the lips divided by the volume velocity at the glottis can be written as (Rabiner and Schafer, 1978)

$$V(z) = \frac{U_L(z)}{U_G(z)}. \tag{2.31}$$

This expression can be expanded by expressing $U_G(z)$ in terms of $U_L(z)$. This is achieved by expanding $U_G(z)$ in terms of the volume velocity equations. The equations of (2.14) and (2.15) give the volume velocity of one section m in terms of the previous section $m-1$. By recursively solving these equations from the first section to the last section a group of equations can be found such that

$$\begin{aligned} \mathbf{U_G} &= \mathbf{Q_G}.\mathbf{Q_{G-1}}...\mathbf{Q_L}\mathbf{U_{L-1}} \\ &= \textstyle\prod_{m=0}^{M-1} \mathbf{Q_m}.\mathbf{U_{L-1}} \end{aligned} \tag{2.32}$$

where

$$\mathbf{U_m} = \begin{bmatrix} U_m^+(z) \\ U_m^-(z) \end{bmatrix} \tag{2.33}$$

$$\mathbf{Q_m} = z^{1/2} \begin{bmatrix} \frac{1}{1+r_m} & \frac{-r_m}{1+r_m} \\ \frac{-r_m z^{-1}}{1+r_m} & \frac{z^{-1}}{1+r_m} \end{bmatrix} \tag{2.34}$$

Note that a fictitious $L-1$ tube (a tube one past the lips) has been added to represent the boundary conditions at the lips in the same manner as all the junctions in the system. This $L-1$ tube is considered infinitely long so that there is no negative-going wave in the tube. This is equivalent to assuming that the $L-1$ tube is terminated in its characteristic impedance and the reflection coefficient at the lips r_L is

$$r_L = \left[\frac{c\rho/A_L - Z_L}{c\rho/A_L + Z_L}\right] \tag{2.35}$$

where Z_L is the lip impedance, and A_L is the cross-sectional area of the tube at the lips.

The equation 2.32 shows that the variables at the input can be expressed in terms of the variables at the output. From Fig. 2.18 it can be seen that the boundary condition at the glottis can be expressed as

$$U_G(z) = \frac{2}{1+r_G} U_{G-1}^+(z) - \frac{2r_G}{1+r_G} U_{G-1}^-(z) \tag{2.36}$$

where r_G is the glottal reflection coefficient and is

$$r_G = \left[\frac{Z_G - \frac{\rho c}{A_G}}{Z_G + \frac{\rho c}{A_G}}\right] \tag{2.37}$$

where Z_G is the glottis impedance, and A_G is the cross-sectional area of the tube at the glottis.

Thus, using 2.32 we can write

$$\frac{U_G(z)}{U_L(z)} = \left[\frac{2}{1+r_G}, -\frac{2r_G}{1+r_G}\right] \prod_{m=0}^{M-1} \mathbf{Q_m} \begin{bmatrix} 1 \\ 0 \end{bmatrix}. \tag{2.38}$$

which is equal to $1/V(z)$.

By multiplying out the equations a general form of $V(z)$ can be found for a lossless tube mode,

$$V(z) = \frac{G}{D(z)} \tag{2.39}$$

where

$$G = 0.5(1+r_G) \prod_{m=0}^{M-1} (1+r_m) z^{-M/2} \tag{2.40}$$

and

$$D(z) = [1, -r_G] \begin{bmatrix} 1 & -r_1 \\ -r_1 z^{-1} & z^{-1} \end{bmatrix} \cdots \begin{bmatrix} 1 & -r_N \\ -r_N z^{-1} & z^{-1} \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \tag{2.41}$$

which can have the form

$$D(z) = 1 - \sum_{m=0}^{M-1} a_m z^{-m} \tag{2.42}$$

such that the transfer function of a lossless tube model has a delay corresponding to the number of sections of the model and it has no zeros, only poles. These poles define the resonances or formants of the lossless tube model (Rabiner and Schafer, 1978).

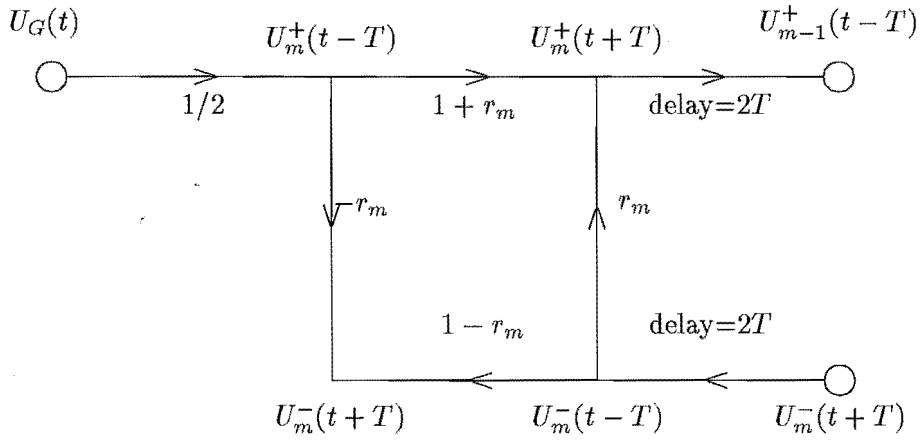


Figure 2.18. Mathematical model showing glottal termination (taken from Markel and Gray(1976))

Chapter 3

SPEECH RECOGNITION - AN HISTORICAL OVERVIEW

..an engineer hoping to build devices that will recognise speech has a right to be discouraged. (Miller, 1962)

Will human beings ever converse easily with computers? Brainwashed by science fiction, films and the media, much of the general public seem to think it is already an everyday event! The reality is of course very different.

Speech, acknowledged as the most effective form of human communication, is also regarded as the most desirable form of information transfer between humans and machines. Thus speech recognition has many worthwhile applications that range from helping the disabled, to simplifying working environments as well as aiding human interaction with machines. It is also one of the many technologies which cannot at present be successfully implemented because the necessary expertise has yet to be acquired.

In the following three sections, which cover the periods before the 1970s, the 1970s, and the 1980s, the history of speech recognition by machines is summarised. In the first section, §3.1, a general overview of the beginnings of speech recognition is given. This history is given as much for its interesting reading as for its background to the modern day speech recognition field. Until the 1970s, research concentrated on the general analysis of speech acoustics and production. During this time basic speech research increased the overall knowledge of speech characteristics and led to the design of some simple speech recognition devices. However, there was little work on problem areas directly related to speech recognition such as timing difficulties, methods of training the systems, selection of features and the effects of noise. Thus, simple recognition schemes never successfully went beyond the laboratory as operating systems.

During the second era, the 1970s, speech research was directed more specifically at the problems of word/speech recognition. It was during this time that many recognition methods were implemented, but in a seemingly random way with no coherent research methodology being adhered to. There were three major problems examined by researchers during this era; the accuracy, the vocabulary and the complexity. Many techniques were suggested to solve problems related to the accuracy of the system. Methods, such as dynamic time warping (DTW) and hidden Markov modelling (HMM) were implemented to reduce the effects of speech timing variations. These methods, known as ignorance methods, were useful because they reflected our ignorance of how speech changes from speaker to speaker, vocabulary to vocabulary and sound to sound, implicitly taking into account the lengthening and shortening of sounds during speaking. Some recognisers used linguistics and syntactics. Grammatical and lexical information was added to some recognisers, to decipher the recognisers' output and hence try to increase their accuracy.

Vocabulary sizes stayed small for both the limited number of commercial recognisers and the many non-commercial recognisers which were presented in the literature. Large vocabulary testing caused difficulties particularly with amount of data collection required to produce large databases and also with the collating and storing of the data

and results.

The complexity of recognition systems hindered the movement of devices from the laboratory to the commercial world. Systems that gave high accuracies and hinted at user generality tended to be highly complicated requiring extensive computer resources. Consequently few systems designed at this time could be exported from laboratory computer to commercial hardware. During the late 1970s some very-large-scale-integrated (VLSI) systems were being designed and the beginnings of digital signal processing (DSP) technology was also hinting at revolutionising recognition systems. The difficulties facing researchers in the 1970s are discussed more thoroughly in §3.2 by examining the two major research areas; isolated word (also known as discrete) recognition in §3.2.1 and continuous (also known as speech) recognition in §3.2.2.

The third section, §3.3, outlines the research during the 1980s and to the present. Although the basic understanding of why one recognition system works better than another is still unknown, the emphasis during the 1980s moved towards establishing which methods may be better than others by introducing standardisation. One important development was the introduction of speech databases which standardised factors such as the number and type of speakers and the noise levels of recording. Standardisation allowed different research teams to test systems and compare results, allowing researchers to establish conditions in which one system may operate better than another. Such experimentation brought researchers to the realisation that there are many variables that can affect the accuracy of recognition systems. The many variables affecting recognition accuracies is one of the problems with speech recognition, and, during the 1980s, many of these variables were examined. The types of variables and methods used are discussed more thoroughly in §3.3 by detailing some of the more important areas of research. The areas of research include; recognition methods (§3.3.1), features tested (§3.3.2), training and testing procedures (§3.3.3), speaker-dependent and speaker-independent recognition (§3.3.4), the vocabularies tested (§3.3.5), and continuous and connected recognition procedures (§3.3.6). For more details of the individual systems of the eras Tables A.1-A.3 in appendix A list many of the systems that were proposed along with the important details of each recognition system.

3.1 THE BACKGROUND TO 1970

The following sections give a historic overview to the beginnings of speech research and in particular details those areas which resulted in speech and word recognition devices. Early speech research began with the modelling of the vocal tract to reproduce speech sounds. Researchers using this modelling process began to realise the link between the vocal tract shape (along with the vocal cords) and the speech produced. Later, when researchers were able to examine visual productions of the spoken words via spectrographic drawings of the speech, links were formed between the spoken word and the drawn word, allowing machines to 'write' the spoken word in the form of spectrograms. Writing the spoken word became the basis of early 'speech-to-text' machines, which later evolved into speech recognisers. The following sections discuss the research in these early years in more detail.

3.1.1 Synthesising Speech.

The beginning of all research into speech and speaking began in the late 19th century with Lord Rayleigh's 'The theory of sound' (1896). This great work is only in part devoted to the analysis of speech characteristics as it concerns all aspects of pressure waves in a gaseous medium. Included in the topics covered by Rayleigh on speech is

the relationship between sound generation by speech mechanisms to that by wind instruments. By relating these forms of sound generation Rayleigh was able to establish relationships between the rate of pulsing of an exciting reed and the pitch of the synthesised vowel. He also investigated the independence of absolute speaking pitch and vowel quality. Another important realisation was that the identity of a vowel depends 'not upon the absolute pitch of one or more resonances, but upon the relative pitch of two or more'. These fundamental discoveries about speech represented the crucial realisation that the same sounds can be made by speakers with vocal apparati of differing sizes and whose voices vary considerably with pitch.

With a better understanding of the production and mechanisms of speech, methods of artificially generating vowels were possible. Rayleigh(1896) noted that a vowel sound is 'merely the rapid repetition of first peculiar note' and that it should follow 'if we can produce this rapid repetition in any other way, we may expect to hear vowels'. The desire to artificially generate speech was reflected by the many papers documenting methods that had been attempted. Rayleigh's work notes the successful methods of Preece and Stroh(1879) and Hermann(1879). Preece and Stroh(1879) synthesised vowels using a rotating disc, similar to a phonograph record, and Hermann's(1879) device consisted of a 'rotating arrangement' with a series of holes, known as a polyphonic siren, producing vowel sounds which were dependent on the number of holes used.

Relating vowel sounds and frequency content, in the form of spectrograms, linked the mathematical representation of sounds and their visual representation. Jones(1918), one of the first to examine the relationship between vowel sounds and frequency content, began classifying and synthesising sounds using frequency analysis. Russell(1929) did much the same as Jones(1918) by documenting the vocal tract shape with the speech frequency content, measuring vocal tract position using X-ray photography.

Dudley(1939) was also interested in the vocal apparatus, and in this case the vocal cords, in relationship to the speech produced. Dudley(1939) noticed the influence of the vocal cords on the type of speech, either voice or voiceless (refer §2.1.2). In experiments to synthesise speech automatically he designed a synthesiser he named a *vocoder*, or voice coder. The vocoder analysed speech and coded it into three quantities; pitch, intensity and frequency spectra. From these three quantities speech could then be synthesised 'almost instantaneously'. Experimenting with the pitch of the synthesised speech Dudley(1939) was also able to relate the affect of pitch to the 'emotional content' of the speech, showing the link between pitch and intonation. During 1940 Dudley exhibited his vocoder publicly, demonstrating its ability to synthesise quality speech.

The visual form of the frequency content of speech, or spectrograms, was believed to hold the key to the information which showed one word as being unique from another. The output spectrograms were known as *visible speech* and were to allow the reading of speech patterns by human and machine. However, the representation of speech as a spectrogram was technologically difficult because of a lack of suitable display devices. Many methods of graphically displaying spectrographic outputs were being studied with the most successful being those proposed by Koneig(1946), Steinberg(1946), Kopp(1946), Riesz(1946) and Dudley(1946). One such system, designed by Kopp(1946) was 'to give the first visible pattern of speech' and be 'primarily of service to the deaf'. However these systems never became widely used outside of the research laboratories where they were invented due to the inability to obtain legible patterns on the output medium that was used - either heat sensitive paper or cathode ray tube screens.

Probably the most important move towards modern day recognition systems occurred during 1948 when speech sounds began to be quantified by parameters, such as formants (refer §2.1.2). The representation of speech as complete spectrograms was far

too difficult at this stage and so during 1948 attempts were made to quantify sounds with particular characteristics. One extremely useful characteristic was to represent speech by its formant frequencies. Potter and Peterson(1948) attempted to categorise vowel sounds by representing them in three dimensions consisting of the first three formants. Further examination of formants led Potter and Peterson(1948) to further categorise sounds by examining the movement of dipthong vowels and the effect of consonants on neighbouring vowel formants, plotting these movements within the three frequency dimensions. Formant differences for different speakers were also investigated by Potter and Peterson(1948), with plots of between-speaker and within-speaker variations.

Speech formants were examined by Dunn(1950) who related formant position with vocal tract position, where vocal tract position was analysed by X-rays during phonation. An approximation to the vocal tract shape using a series of cylindrical sections was produced for a set of vowels. For each vowel shape Dunn(1950) calculated the multiple resonances produced by the cylindrical approximation using distributed impedance theory, representing the cylinders by an equivalent electrical network. Calculating the three formant positions for a series of vowel sounds, Dunn found that for each vowel the formants fell within a predicted range of the calculated values. Using this equivalent circuit with a suitable excitation, Dunn was then able to synthesise vowel sounds.

Potter and Stienberg(1950) initiated a rigorous study to label sounds with characteristic properties. Taking many occurrences of a particular sound, they looked for repeated frequency events that would classify the sound. Although they obtained high selectivity, with this method, for a single speaker, they did point out that it would be difficult to classify the vowel sounds across a range of different speakers.

Chang, Pihl and Essigmann(1951) chose to represent sounds by more than just frequency using autocorrelation and infinite clipping (also known as zero crossing rate refer §4.2) as well. They conjectured that these parameters describe some 'essential elements of speech sounds that are statistically invariant'. However they concluded that speech sounds are far too complicated to be accurately represented by only these types of mathematical parameters.

By the early 1950s speech analysis was at a point where speech sounds could be analysed in frequency and time. Using frequency analysis, similarities between different utterances of the same sound had been observed. However it was obvious there existed problems with variations from speaker to speaker.

3.1.2 Mechanical Speech Recognition

In the early 1950s primitive recognition devices were being built which used amplitude versus frequency information.

The first electronic recogniser has been attributed to Jean Dreyfus-Graf(1950) (Lindgren, 1965a) of Geneva Switzerland. Considerably less sophisticated than other methods of the 1950s, Dreyfus-Graf's recogniser did not in fact recognise speech but simply translated speech into a form of written word. Dreyfus-Graf spent many years researching the design for which he used six filters to 'divide down' the acoustic spectrum 'to the six principal formants of the mouth orchestra'. The six filters operated a pen recorder providing diagrams of the input sounds. An illustration of Dreyfus-Graf's 'recogniser' is drawn in Fig. 3.1. Dreyfus-Graf believed these 'sound drawings' to be equivalent to prealphabetic Chinese or Egyptian pictographic signs hence allowing the machine to 'write directly what we say'.

A more modern method was proposed by Smith(1951) which actually did distinguish speech sounds. Smith's impressive attempt to recognise sub-word sounds, in this case phonemes (refer §2.1.3), used a number of cathode ray tubes in parallel providing

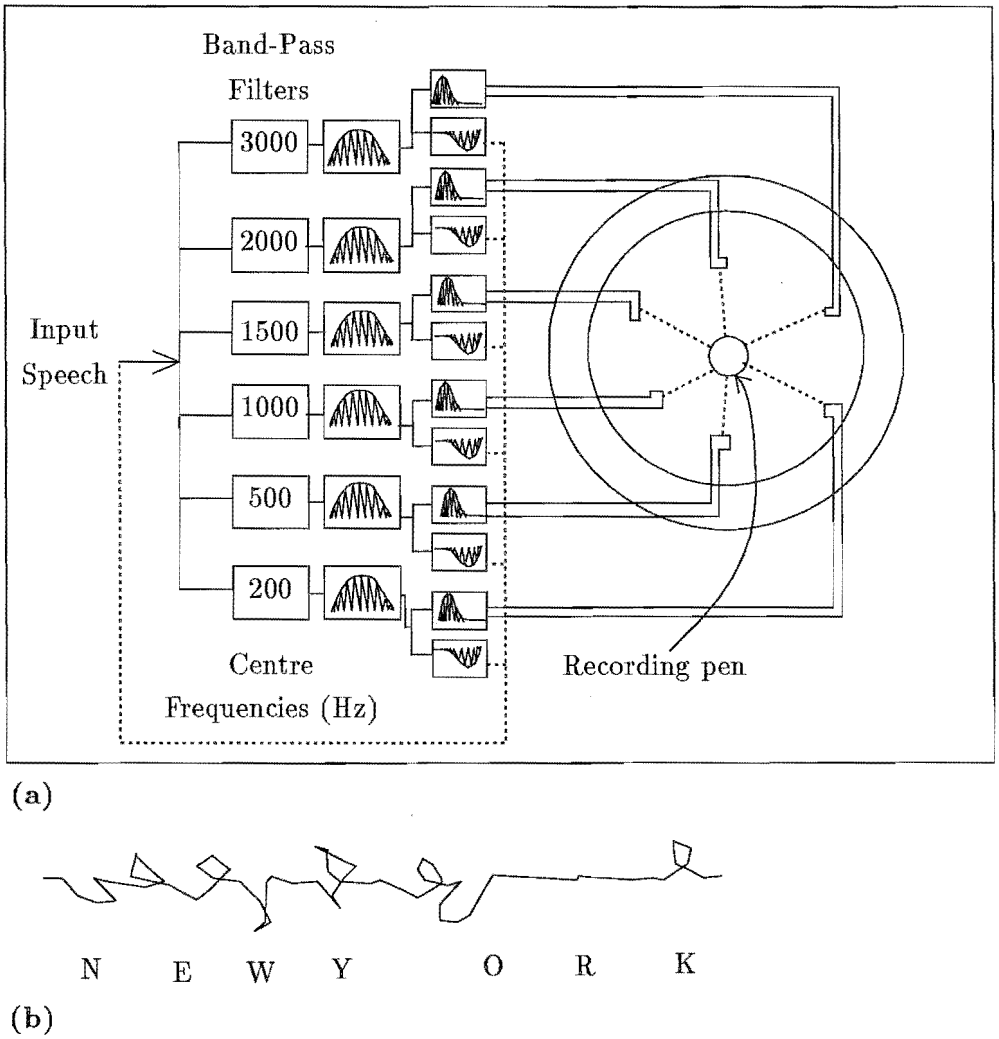


Figure 3.1. The first speech-to-text machine invented by Dreyfus-Graf(1950). The illustration in (a) shows Dreyfus-Graf's 'recogniser'. Six filters were used to 'divide down' the speech into the 'six principal formants of the mouth orchestra'. The machine wrote the words spoken in a 'sonographic' alphabet shown in (b).

a simultaneous display of visible speech frequency patterns from a common speech signal, shown in Fig. 3.2. A mask, representing one of the sought-after (or reference) phonemes, was superimposed over each visible speech image. The light transmitted from each cathode-ray tube through its mask was measured using a photocell. No results were reported for this method and no further references to this method were made, probably because a number of superior recognition methods were soon to be proposed. However this method typifies many later attempts at speech recognition where a comparison is made between a test pattern and a series of reference patterns, and a recognised word is chosen based on a distance measure.

Using the same basic method as Smith's(1951) the first high-quality recogniser can be attributed to Davis, Biddultph and Balashek(1952) from Bell laboratory. They acknowledged the difficulty of recognizing sounds due to the variability encountered in repeated utterances of the same word by different speakers and so were the first to develop their recognition technique as a speaker-dependent one. Their system, as with Smith's system, was based on a pattern matching technique, which compared

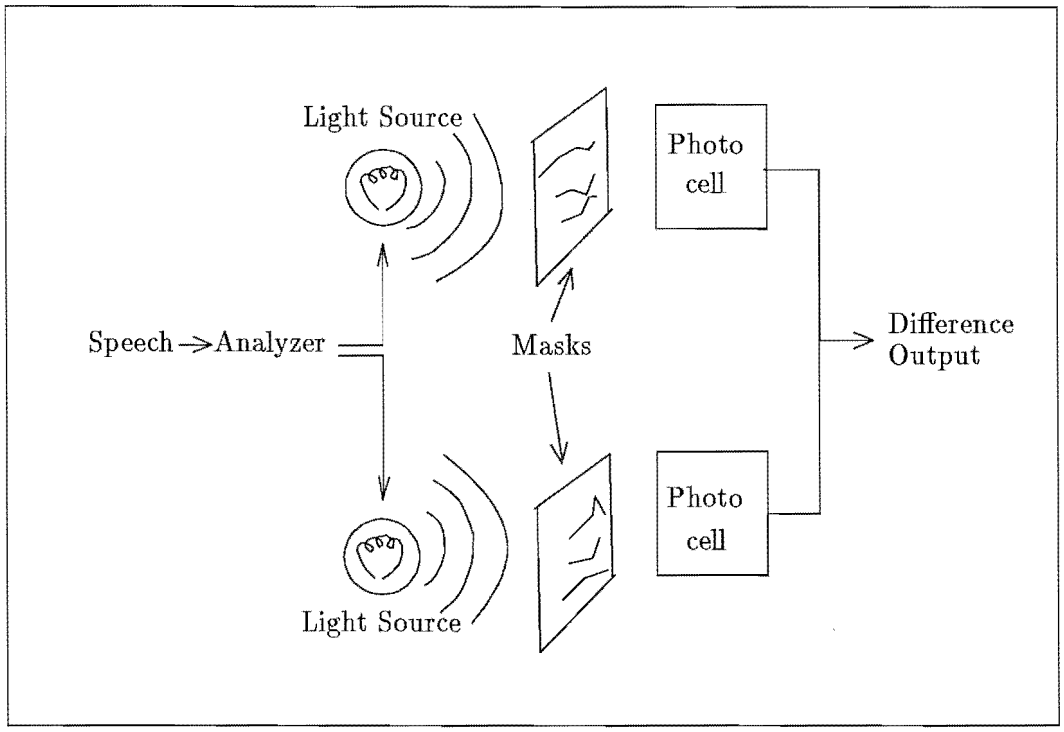


Figure 3.2. Smith's(1951) attempt at recognizing sounds consisted of measuring the light projected through masks of the speech sounds. The masks of the sounds were made from spectrograms of each sound.

aspects of data derived from an unknown signal (*test* template) with the corresponding aspects of a number of known signals (*reference* templates). A block diagram of the recogniser circuit is illustrated in Fig. 3.3. The vocabulary tested was the digits ZERO to NINE. Reference templates of these words were formed during the training phase by storing the first two formants of each word whenever a significant change to the formants occurred. A recognition rate as high as 98% was achieved. When tested in a speaker-independent context the accuracy dropped to around 50% and the authors noted the importance of tuning the recogniser to the speaker.

From these tentative beginnings many researchers began realising the potential of word recognition systems. A recogniser which used a unique method of sound recognition was described by Wiren and Stubbs(1952) and is illustrated in Fig. 3.4. Based on frequency analysis, the system used a method of binary classifications which continued until a sound was fully described. The categories, based on frequency analysis, required decisions such as turbulence/nonturbulence, acuteness/graveness, compact/diffuse, nasal/nonnasal and stop/fricative. Testing on vowel sounds produced an accuracy of 85 to 94% depending on the vowel. Olson and Belar(1956), from RCA laboratories, implemented the first successful voice controlled machine, in this case a typewriter. Control was achieved by mechanically recognizing syllables of sounds. The typewriter produced the speech in written form as phonetically based text.

Dudley and Balashek(1958) and Fry and Denes(1957,1958) also produced functioning mechanical recognition systems based on the phonetics of speech. Dudley and Balashek continued the Bell laboratory system design, discussed previously, of a mechanical recogniser for digits. Their method, however, recognised the phonetic patterns of speech. Under laboratory conditions, for a single speaker, the authors claimed 'al-

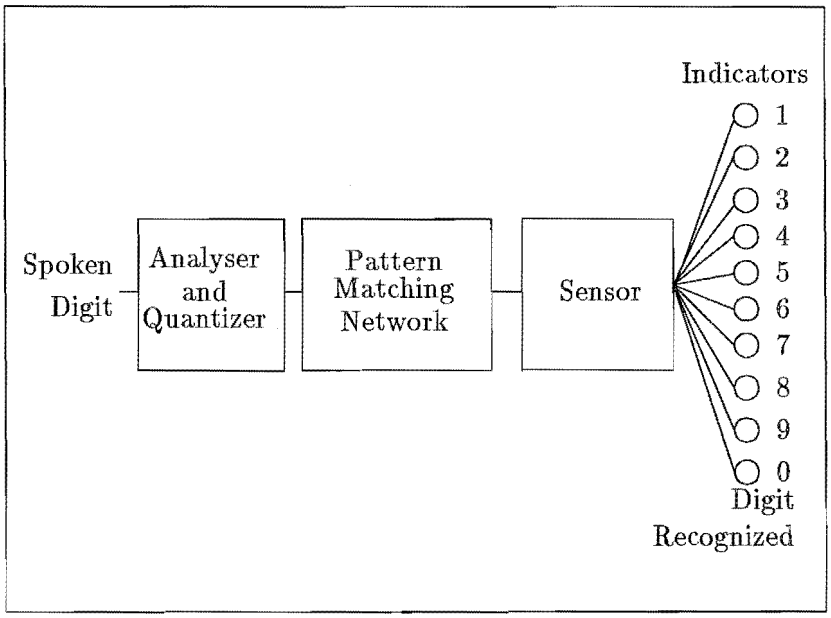


Figure 3.3. A block diagram illustrating Davis, Biddultph and Balashek's recogniser(1950), a speaker-dependent recogniser for the digit words ZERO through NINE. The recogniser is trained for each digit word, then, by means of a comparison between a test and each reference word, the digit of best match is selected.

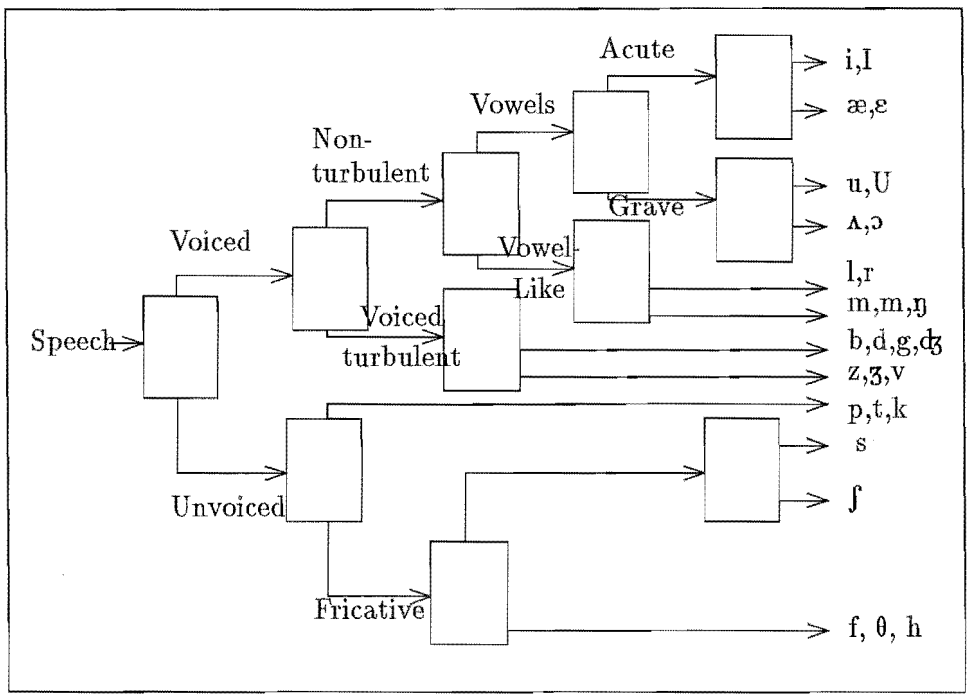


Figure 3.4. The first mechanical recogniser to recognise phoneme sounds by successive binary classification. The first step separates voiced from unvoiced, then turbulent from nonturbulent. Nonturbulent sounds are classified into 6 groups. Unvoiced turbulent sounds are classified into stops and fricatives. Only the vowel sounds were completely distinguished.

most perfect recognition'. They also noted that with different pronunciation or with different speakers a 'considerable number of errors occurred', however, if the speaker 'corrected' his or her voice, accuracies as high as 90% for speaker-independent recognition could be achieved.

Fry and Denes' mechanical speech recognition system also recognised isolated phonemic patterns of speech leading to the recognition of complete sentences. This method recognised phoneme units which were firstly built into words and then into sentences. Accuracy results of 60% for phoneme recognition and 24% correct for word recognition were quoted, and although accuracies are low this method was far advanced for its time. Fry and Denes discussed the problem of training the system for the speaker. They suggested the system be trained for each new speaker by using a feedback scheme. This scheme would require a known sentence to be spoken and the system adjusted to the individuals' frequencies by correcting any errors made in the training stage.

A method of successive discrimination was used by Forgie and Forgie(1959) who attempted vowel recognition based on frequency analysis obtaining a recognition accuracy of 93%. Their method initially gave a coarse quantised frequency estimate of the first two formants. The initial estimates were very general so that up to as many as six vowels could have the same F1/F2 quantised location (only two vowel phonemes could be distinguished uniquely at this point), see Fig. 3.5. Discrimination was achieved by adding finer detail to the formant structure information such as formant bandwidth. The significant difference in this system from previous recognition systems was that it could tailor the particular information used to discriminate between the vowels based on the initial coarse formant calculations.

By the early 1960 many promising methods of speech recognition had been tested, however the only systems that had shown any true promise, giving high accuracies, were those that recognised single sound elements, spoken carefully by selected speakers and set apart with silent pauses. The leap from this point into a real life situation, which had seemed 'around the corner' (Lindgren, 1965a) was now very unlikely. One viewpoint, from Bell Labs was 'we should consider it unfair of us to expect so much of the machine. Perhaps the interface between man and machine must be set at some other point to demand less of the machine' (Lindgren, 1965a).

3.1.3 Timing Variations

Accuracies of recognition systems were being limited by the problem of word length variations. Comparisons of words with different lengths could not be effectively achieved unless the word lengths could be first normalised. During 1960 both Denes and Mathews(1960) and Olson and Belar(1961) noted problems with talking speed variations and hence word length variations, on the frequency patterns of speech. Olson and Belar(1961), showed how the frequency pattern of the same vowel, spoken at different rates, looked different when analysed using fixed length time intervals. They compensated for the timing effect by recording a spectrum whenever a significant change of frequency was noted. Although the paper gave no results, it did point out that memory saving of greater than 50% could be achieved if fewer template patterns were needed as a result of removing the time variations.

Denes and Mathews(1960) also introduced time normalisation of time-frequency patterns for their recognition task. The method of time normalisation of the frequency pattern they describe as visualising the pattern to be 'painted on a rubber sheet, the time normalisation consists of stretching the sheet until the distance between the beginning and end is a standard length'. This was the beginnings of the idea of linearly distorting one pattern of speech to fit another. The recognition results for time normalised patterns were better than non-normalised patterns, with errors of 6 and 13%

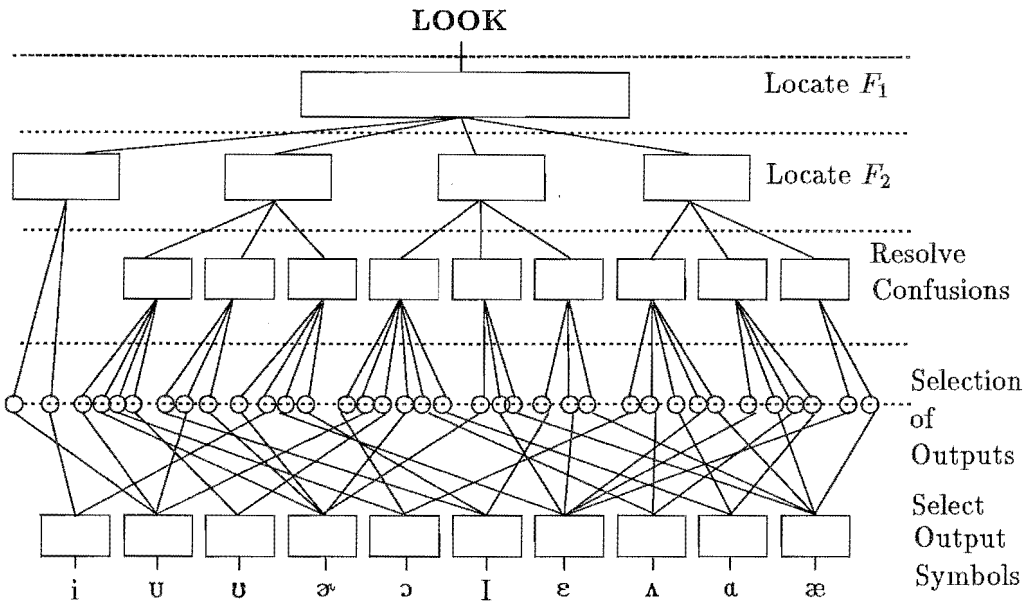


Figure 3.5. The Forgie and Forgie(1959) vowel recogniser. This recogniser classified each 'LOOK' at a sound into 10 possible vowel classes. Formant structure was used to initially coarsely quantise the sounds, then further information was added to resolve any vowel confusions.

respectively.

In 1968 Shearme and Leach(1968) pointed out that 'the most serious problem encountered in spectrum matching is the well-known lack of synchronism between corresponding spectral events when phonemically identical words are spoken by different talkers'. They devised a time normalisation method for their spectrum recognition scheme. Based on work previously by Olson and Belar(1960), quantised spectra were produced by saving the spectrum only when a significant change occurred, thus reducing variations in the spectrum with respect to speech timing. Each set of spectra for a word was normalised to a set length by the addition of silence. In tests with ten speakers an accuracy of 90% was obtained with a 32 word vocabulary. However no comparison without this time alignment was given. Shearme and Leach(1968) also produced a speaker-independent scheme in which many templates taken from a selection of speakers were used. Using multiple templates was believed to be the best way of achieving accurate speaker-independent recognition, taking into account the effects of speaker variations.

Reddy(1969) also reported problems with aligning speech segments across time. His method of recognition matched the segmental parameters of the utterance being tested with the corresponding segmental parameters of the known reference utterance. This was achieved by initially labelling the segments across each word as belonging to particular phoneme group and during the synchronisation procedure mapping vowel to vowel and fricative to fricative. He reported 98% success for a 50 word vocabulary.

3.1.4 Continuous speech recognition

The design of real-time continuous recognisers was being seriously considered in the late 1960s. Although the ideas behind continuous recognition had been discussed for some time the implementation was considered merely a dream. In 1961 Marrell reviewed the

speech recognition technologies and noted the ultimate goal was to produce 'a device which, in effect, replaces a secretary taking dictation', he also added 'the achievements to date have been modest', with 'the best that has been done so far is the recognition of the spoken digits and of some of the vowels in isolated context'.

There were, however, some relatively successful attempts already trialed in this area. Olson and Belar(1956) first described a method which recognised a vocabulary of ten words, and formed these into sentences as the user spoke. This method was designed into their 'phonetic typewriter' in the early 1960s.

One way to conquer the continuous recognition problem was proposed by Halle and Stevens(1962) and consisted of introducing morphological and syntactic rules into the basic recognition algorithm to take advantage of the structure of language. Yet another method, discussed by Sakai and Doshita(1963) was for a 'conversational speech recognition system'. The system operated by segmenting the speech into phonetic intervals which could then be recognised. The segmentation divided vowel from consonant and vowel from vowel using two quantities that Sakai and Doshita call *stability* and *distance*. Stability measured the stationary property of a pattern while distance measured the change. These segments were classified into either consonant or vowel by using a measure of zero-crossings (refer §4.2). Testing on Japanese male speakers saying monosyllables, which contain a consonant and a vowel, Sakai and Doshita(1963) obtained an accuracy of 90% for the vowel segment and 70% for the consonant segment. Sakai and Doshita noted that for high recognition accuracies the input word must be clearly and slowly spoken.

Although work continued on continuous recognition, particularly in the area of segmenting words into phonemic units (Reddy, 1966), (Reddy, 1967a), a computer based scheme did not appear until 1967 (Reddy, 1967b). Reddy discussed a scheme which produced a phonemic transcription from a connected speech sample. This system segmented the speech into stable intervals, where the parameters remain in an almost constant state, and transient intervals, where parameters undergo gradual change. Pitch synchronous spectral analysis was used for classifying these states into phonemes. Results from classifying 32 continuous speech utterances, or 287 phonemes, indicated 81% correct phoneme recognition. Reddy concluded by stating that 'connected word recognition is not now a utopian dream, but indeed a distinct possibility'.

3.1.5 Linguistics

With the slow progress of recognition machines based on acoustic information, many speech recognition researchers were now moving away from pure acoustic analysis. Other methods of analysing speech were being examined. One method was to find the process which humans use to decipher speech. This research showed the importance of linguistic cues for human speech recognition. One paper on this topic (Lindgren, 1965b) pointed out 'we must at last confront the problem of language...if engineers are to take seriously the idea of building automatic speech recognition machines, they can no longer avoid the questions related to language itself.'

Denes(1959) also believed linguistics to be important. In his study of the design and operation of his mechanical word recogniser he wrote with great insight that he believed 'no simple relationships exist between the spectral patterns and speech sound units' and 'it seems likely that the human listener does not rely solely on acoustic characteristic and it is unlikely that any single acoustic characteristic, or combination of characteristics, uniquely identifying any speech sound does in fact exist in the acoustic wave'. Denes developed a method of recognition which used a basic acoustic recogniser with its output modified by a set of linguistic cues. His scheme had the initial acoustic recognition acting as a phoneme recogniser via a spectrographic representation. The

recogniser, however, did not recognise much better than a purely acoustic recogniser with only a 72% accuracy for thirteen phonemes, four vowels and nine consonants and a 40% accuracy for the recognition of complete words

3.1.6 Commercial Interests

Now that some isolated recognition accuracies were reaching acceptable levels (>90%), commercial applications were becoming increasingly evident. Systems that could be produced with realistic commercial attributes - low cost, accurate and portable were required. However, systems were still firmly centred around research laboratory computers, and although reasonably accurate in that environment could not be implemented in a form that could be marketed. Many problems associated with commercial system such as the effects of noise, transmission line variations, telephone lines, microphones, different users and different environments, had not yet been considered.

Some researchers were trying to overcome the problems of taking a recognition systems out of the laboratory. Teacher *et al*(1967), expressed concern that no development groups were taking word recognition equipment from the 'academic curiosity' stage into the field of commercialism as products. They designed a commercially viable system consisting of a portable mechanical word recognition machine. They based their scheme on three speech features - a single equivalent formant frequency, calculated from the first three formants of speech, the equivalent formant amplitude and the *state of voicing*, (voice/unvoice, refer §2.1.2). The recogniser was tuned for a specific vocabulary, the numerals OH through NINE. Tests gave 90% accuracy, and 1% misrecognition.

Although there were few commercial systems available there did appear to be many systems discussed in the literature that would have been ideal for commercial usage. A recognition scheme described by Gilli(1967) was similar to that of Teacher *et al*(1967). It was designed to recognise the ten digits for many speakers. However Gilli's system was designed on a digital computer. The power spectrum of the sounds was used with results of 90% success quoted for trained speakers and 70% success for untrained speakers.

Another scheme that could have had commercial success was that of Ewing *et al*(1968) who produced an accurate, speaker-dependent scheme using zero-crossing information (refer §4.2). Based on a computer the system recognised the vocabulary ZERO through NINE with 'excellent' results, although they also noted their results were not dependable for all speakers.

Purton(1968) tried using autocorrelation analysis for his recognition scheme claiming this method was as attractive as frequency analysis with less computation and easily implemented in hardware. Testing this method on the vocabulary NOUGHT through NINE, the accuracy for a single speakers ranged from 78 to 99%. For speaker-independent tests accuracy dropped to 59%.

The production of commercial systems was looking promising with a limited vocabulary, limited speaker system. However, for these systems to obtain accurate recognition (>90%) complicated processing was required. Purton(1968) believed the two main hindrances limiting the commercial application of speech recognition was the relatively long recognition time and the large computer storage requirements. Purton's scheme required 3 seconds per utterance for the recognition of his 10-word vocabulary.

3.1.7 Summary

Even though many high quality recognition systems were being reported, see Table 3.1, the main requirements for a commercial automatic word recognition system - speed and generality of user - had not been met. Lavington(1969) represented the feelings

among most of the researchers in the field by being sceptical about whether many of the basic areas of speech recognition had yet been examined properly, especially the relative merits of many of the processing techniques used for recognition. These feelings were due to the multitude of schemes testing individual methods with very little cohesion or comparison between each system. This random, uncoordinated testing paradigm would continue to worsen into the 1970s.

3.2 THE 1970S

The first models [for speech recognition] failed to recognize speech successfully showing that our theories were inadequate. As the years passed, our ideas about speech recognition - and the models built to implement them - became more and more sophisticated, and yet, the results are still unsatisfactory. (Denes, 1964)

Denes' quote of 1964 could be easily applied to the research undertaken in the 1970s. Speech recognition in the 1970s broke away from the general research of speech analysis. This surge of interest towards speech recognition was sparked by both the increased use of computers and the interest from the commercial field. By concentrating solely on word and speech recognition many of the problems inherent in these methods such as segmentation, classification, timing, time normalisation, pattern matching, and training methods such as clustering, were thoroughly examined. However, many general speech problems such as the limitations of particular speech feature representations and general signal processing techniques were ignored. Technological advances in the 1970s allowed the implementation of more complicated and sophisticated methods with very little improvement of results. Computers eased many of the speech recognition problems allowing the storage of large databases, repetition of testing, and collating of results required for examining recognition problems.

Isolated word recognition (also known as discrete recognition) systems had been researched for over a decade and many believed that discrete recognition had been thoroughly examined. This was hardly the case however, as there were still problems with achieving a reliable, fast system for multiple speakers. Many discrete recognition systems attempted in the early 70s continued in the vein of those recognisers of the 50's and 60's, basing recognition on spectral calculations and phoneme classification. One problem stopping these recognisers becoming commercially successful was that although these devices worked successfully in demonstrations they did not maintain performance levels in practical applications (Carpenter and Lavington, 1973).

During the mid 1970s discrete recognition systems would again be studied with the introduction of time normalisation techniques such as dynamic programming (DTW) (refer §5.1) and to a lesser extent hidden Markov modelling (HMM) (refer §5.2). These methods would revolutionise discrete systems, and also other areas of recognition research that relied on discrete recognition, such as connected word recognition.

Research into continuous recognition and speech understanding systems boomed in the 1970s. Many who had begun researching discrete recognition systems were now rushing into this new field of research. One example of the popularity of continuous recognition is given by the attendance of a continuous speech recognition symposium, held by the IEEE, in 1974. The symposium hosted over 130 researchers interested in the field of continuous recognition.

Because, in the 1970s, recognition systems seemed to break into two major areas, either word or speech recognition, these two areas will be the topics discussed in more detail in the following sections, §3.2.1 and §3.2.2 respectively.

3.2.1 Discrete Word Recognition

The goals of discrete recognition systems were that they operate in real-time, be used by multiple speakers and give high recognition accuracies. Obtaining these goals would give the discrete recognisers attributes that would make them highly attractive to commercial interests.

3.2.1.1 The Discrete Systems

Simple discrete recognition systems were able to take advantage of fast computer hardware such that real-time operation was possible, although for more complicated systems times of up to 16 seconds for recognition of a word were quoted (Clark, 1970; Beninghof and Ross, 1970). One system, produced by Scarr(1970), and based on a non-alignment method, clearly used the speed advantages that a non-alignment scheme (explained below) gives, producing a speaker-independent digit recognizer obtaining 95% accuracy with male speakers in real-time. Scarr concluded 'we are still a very long way from the solution to the problem of continuous speech and large vocabulary. The problem of isolated words and limited vocabulary is tractable'.

Many systems, like Scarrs', were tractable, using a multitude of simply calculable temporal and spectral features (Clapper, 1971; VonKellar, 1971; Rabiner and Schafer, 1978). From these acoustic representations the systems went on to produce a string of sound classes, usually as a set of phonemes (Itahashi *et al.*, 1973; Scarr, 1970). However, some systems classified the sounds into more unique classes, such as nasal-like, and fricative-like, or front vowel and back vowel, or silence, quasi-stationary, non-stationary, and transient (VonKellar, 1971; DeMori, 1973; Sambur and Rabiner, 1975). Time normalisation was, therefore, not required as each word was recognised by its sequence of sound classes. Final classification was achieved by examining lexicon knowledge (Scarr, 1970).

Further problems existed with speaker-independent recognition. Some researchers tried to solve the problem of speaker-independent recognition, as in the 50s and 60s, by using multiple templates from many speakers. Another method, attempted by Miller(1970) was to remove the problems of vocal variations between speakers by using a learning (also known as an adapting) recognition system. Adaption occurred with the system compensating for formant changes. His design was to operate in real-time ('less than a few seconds') on the ten word vocabulary ZERO through NINE giving speaker-independent recognition rates of under 90%.

Although many methods were tested for discrete recognition little improvement was being made. Itahashi(1973) wrote of his discrete recognition system that 'in spite of the fact that so many research projects have been conducted, no entirely acceptable speech recognizer has been developed'. This pessimism could be attributed to the fact that although much work had been carried out no coherent attempts had been made to tackle the difficulties of noise, environment, and speaker variation (intrinsic to speech and word recognition) and hence few systems could run successfully outside laboratory conditions where these variations are minimal. Pierce(1969) was one of the first to criticise the apparent lack of recognition improvement in his scathing letter entitled 'Whither speech recognition'. The stagnation of the recognition area was again brought up by Pols(1971). Pols(1971) noted that although many schemes had been tested the major stagnation in this field was caused by the lack of research centered on producing an indepth study of the techniques used. He noted that different systems would differ markedly in the features used with no reasons ever given by their authors for choosing a particular feature or a particular system. Moore(1977) also noted these same problems, emphasising the lack of suitable standards for evaluating a recognition

system's performance. Moore pointed out that the minimum requirement to produce a comparison between different system performances would be that they share 'at least' the same vocabulary and the same acoustic samples. However very little notice was taken of these criticisms and system testing was far from standardised.

3.2.1.2 Time Normalisation

The introduction of time normalisation methods, such as dynamic time warping (DTW) (refer §5.1), totally revolutionised word recognition systems. DTW, first discussed in 1960 by Denes and Mathews, was again used in 1970 by Sakoe and Chiba, and also Velichiko and Zagoruik(1970), . Although the technique proved to be more successful than previous methods ('Using the technique of dynamic programming remarkably high scores are obtained' (Fujimura, 1975)) its disadvantage was that it required large computational power to be implemented.

Velichko and Zagoruik(1970) were the first to implement a recognition system using dynamic time warping. Velichko and Zagoruik's scheme implemented a very general dynamic programming algorithm, recognizing a large Russian vocabulary. Requiring 'a great deal of computer time' the overall recognition rate was 95%.

A more constrained dynamic programming (DTW) system discussed by Itakura(1974) claimed very high recognition rates. This system recognised a 200 word vocabulary for a single speaker using a dynamic programming method. An accuracy of 97.3% for telephone quality speech was claimed. One significant improvement by Itakura was the introduction of a new method of distance calculation for all-pole model representation of speech. This distance measure (discussed in §6.4) was also perceptually apt as it involved greater weighting of spectral poles than spectral nulls in the distance measure.

The single most useful advantage of DTW was that it allowed the comparison of words using a standard recognition algorithm. Thus DTW allowed recognition schemes matching procedures to be standardised and hence allowed for the comparison of many parameters used for recognition. One of the first comparisons of the abilities of different parameters using dynamic programming matching was undertaken by Ichikawa(1973). Ichikawa(1973) tested spectrum, cepstrum, autocorrelation coefficients, LPCs and partial autocorrelation coefficients with respect to the recognition of spoken digits (in Japanese). Ichikawa(1973) believed that one of the 'most difficult obstacles' had been removed with the introduction of nonuniform time pattern matching and that the next thing was to 'solve the problem of determining which is the best speech parameter'.

Another comparison of techniques using controlled experimentation was undertaken by White and Neely(1976). They noted that this was one of the first published comparative studies of its kind combining techniques of DTW and linear time normalisation with LPC and bandpass filtering. Tests on the alphanumeric (AN) vocabulary by a single male speaker and also on 91 names by one male speaker showed dynamic processing as superior. Comparisons with results by Itakura(1974) were discussed although very few similarities existed between the two systems except the recognition method used.

DTW was also being used for connected speech recognition. Haton(1974) applied a real-time isolated-word recognition system to the recognition of sentences spoken word-by-word. The acoustic level matching used a DTW algorithm. A knowledge based scheme was also used to predict the incoming word using syntactics and semantics before acoustic recognition, hence reducing the comparison reference set at each word match. The recognition scheme, using a constrained lexicon of 36 possible French words and a fixed syntax, was used to control a machine tool. An accuracy of 99.8% was given. The results show the usefulness of syntactic and semantic constraints working with an acoustic matching scheme for connected speech.

Sakoe(1979) published a connected speech recognition scheme using dynamic programming matching. Called two-level DP matching this system matched words and then phrases using DTW. 100% accuracy was quoted when testing on 200 sentences composed of 3-digit continuous numerals which had been spoken by one male speaker. For five speakers, in speaker-dependent mode and testing up to 4 digit sentences an accuracy of 99.6% was achieved.

Dynamic programming was proving to be an extremely useful, but highly computationally time consuming method. Many faster implementations were being found. Its usefulness and generality along with its high accuracy would encourage development of recognition schemes using this method to continue into the 80s and 90s.

3.2.2 Recognizing Speech

Most papers on speech recognition conclude by saying that it is necessary to use higher level linguistic cues to obtain acceptable recognition. The terms context, syntax, semantics, and phonological rules are used but attempt to utilize these sources of knowledge have not been successful because of the ill structedness of these concepts. (Reddy *et al.*, 1973)

The goal of speech recognition system research in the 1970s was to produce a system that allowed continuous verbal communication between man and machine. The ultimate aim would be the ability to converse with a machine as if speaking to a human being. This Utopian dream had its first practical implementation in the early 1970s.

In the early 1970s this desire was fuelled by grants from the United States Advanced Research Projects Agency (ARPA) which saw more than 20 functioning speech recognition and speech understanding systems constructed in a seemingly coherent and systematic manner. However, many areas, such as vocabulary, environment, and speaker were not specified, and so from the large beginning only three systems were to come close to a useable product which satisfied ARPA. These three systems continued to be funded over the next five years and were the Bolt Beranek and Newmann Inc (BBN) Speechlis system, the Carnegie-Mellon University (CMU) HARPY system and the System Development Corporation system jointly developed with the Stanford Research Institute (SDC/SRI). Many other systems, initiated by this study, also showed extreme promise and continued to be developed with their own funding. These systems were the CMU HearsayII, the BNN hear what I mean (HWIM) system, the Lincoln Laboratories system, the IBM speech understanding system, and the Baker DRAGON system.

The recognition systems from the ARPA project followed strict rules on recognition standards, as they were designed from the same criteria, however there were major differences with the most important being outlined in Table A.2, of appendix A. The criteria forced the systems to recognize speech by deriving the meaning of the spoken phrase rather than using the conventional method of recognizing the absolute word content, word-by-word. Deriving speech meaning forced systems to use multiple sources of knowledge, such as lexical, grammatical, syntactical and semantical information, working together in an effective way and going beyond earlier systems' capabilities (Medress, 1978).

The systems in the ARPA project usually began their recognition by collecting acoustical data from the speech signal in the 0-5kHz range(Woods, 1975; Ritea, 1974) although some went as high as 10 kHz (Walker, 1974). A multitude of parameters were calculated from the data. Combinations of LPCs, signal energy, pre-emphasised signal energy, energy in frequency bands, autocorrelation coefficients, normalised LPC error, (Woods, 1975), zero-crossing counts, LPC derived vocal-tract parameters, 1/3 octave filter band energies (Lesser *et al.*, 1974; Ritea, 1974) or output of digital filter bands

(Walker, 1974) and fundamental frequency (Lea *et al.*, 1974) were used. Generally initial parameters were used to segment the speech into broad classifications using stable characterisations of the signal. Such broad classes included voiced/unvoiced, or phoneme like representations that the author invented such as vowel-like, fricative-like. The segmentation was inherently troublesome and difficult. One researcher noted 'it is not the segmentation scheme that is important but the error correcting scheme used after segmentation' (Reddy, 1976). By using more elaborate techniques these initial classifications were split into more highly defined classes such as phonemes (Niederjohn and Thomas, 1973), diphones or syllables (Ritea, 1974; Walker, 1974). Structural information involving semantics, syntactics and lexicon knowledge was then used to constrain the recognised sentence to a known structure (Barnett, 1973). In this way errors that occurred at the word recognition level could later be corrected at the phrase recognition level.

This advantage of multiple sources of knowledge can be observed in the recognition scheme of Bolt Bernek and Newman(BBN) where the Speechlis system was developed. W.A. Woods along with many associates (Makhoul, Schwartz, Wolf, Scandora, Colarusso, Klatt, Cook, Rovner, Nash-Webber, Bates, and Gould) designed the BBN continuous speech understanding scheme making use of techniques of artificial intelligence, natural language processing and acoustical and phonological analysis and signal processing. They also noted the importance of pitch, energy and segment duration which are present in the spoken utterance and which relate the speech signal directly to the syntactic structure of the utterance. Along the same lines as the Speechlis system, HearsayII, a second HEARSAY system, was being developed by Lesser *et al*(1974) and Erman(1977). HearsayII speech understanding system (Carnegie-Mellon University) was derived from two years study of the original Hearsay system (Reddy *et al*(1973)) and was based on the view that connected speech processing can only be handled through the efficient use of multiple, diverse sources of knowledge. HearsayII's goal was to provide a multiprocess-oriented software architecture to serve as a basis for systems of co-operating (but independent and asynchronous) data-directed knowledge-source processing. The purpose of such a structure was to achieve effective parallel search over a general artificial intelligence problem-solving graph, employing the hypothesise-and-test paradigm to generate the search graph and using a uniform, interconnected, multilevel global data base as the primary means of interprocess communication. An illustration of this system and the details of the process is shown in Fig. 3.6. No results were given at this time, however 90% sentence accuracy was later quoted by Erman(1977) .

All the systems so far discussed used a *bottom-up* approach to recognition, reducing the speech to an acoustic level from which parameters are derived that represent the acoustics of the speech. From these parameters phonemes are recognised, and then words and finally sentences. A bottom up approach was not the only way used to recognize sentences and phrases. Some recognition schemes took advantage of a *top down* approach, using prosodic information to derive phrase and sentence structure first (Lea *et al.*, 1974; Miller, 1962). Examples given by Lea *et al*(1973,1974) showed that recognition schemes that rely upon simple concatenation of categorised phonetic segments could never be completely successful. Preliminary prosodic analysis allowed sentence nucleus positions to be found and recognition to be achieved around this point. Recognition of highly stressed sounds tended to give higher word recognition and segmentation accuracy (Lea, 1973; Lea *et al.*, 1974). Lea(1974) used the idea of applying prosodic features in his analysis-by-synthesis recognition application illustrated in Fig. 3.7.

Also working with a top-down approach was Lindblom and Svensson(1973). who

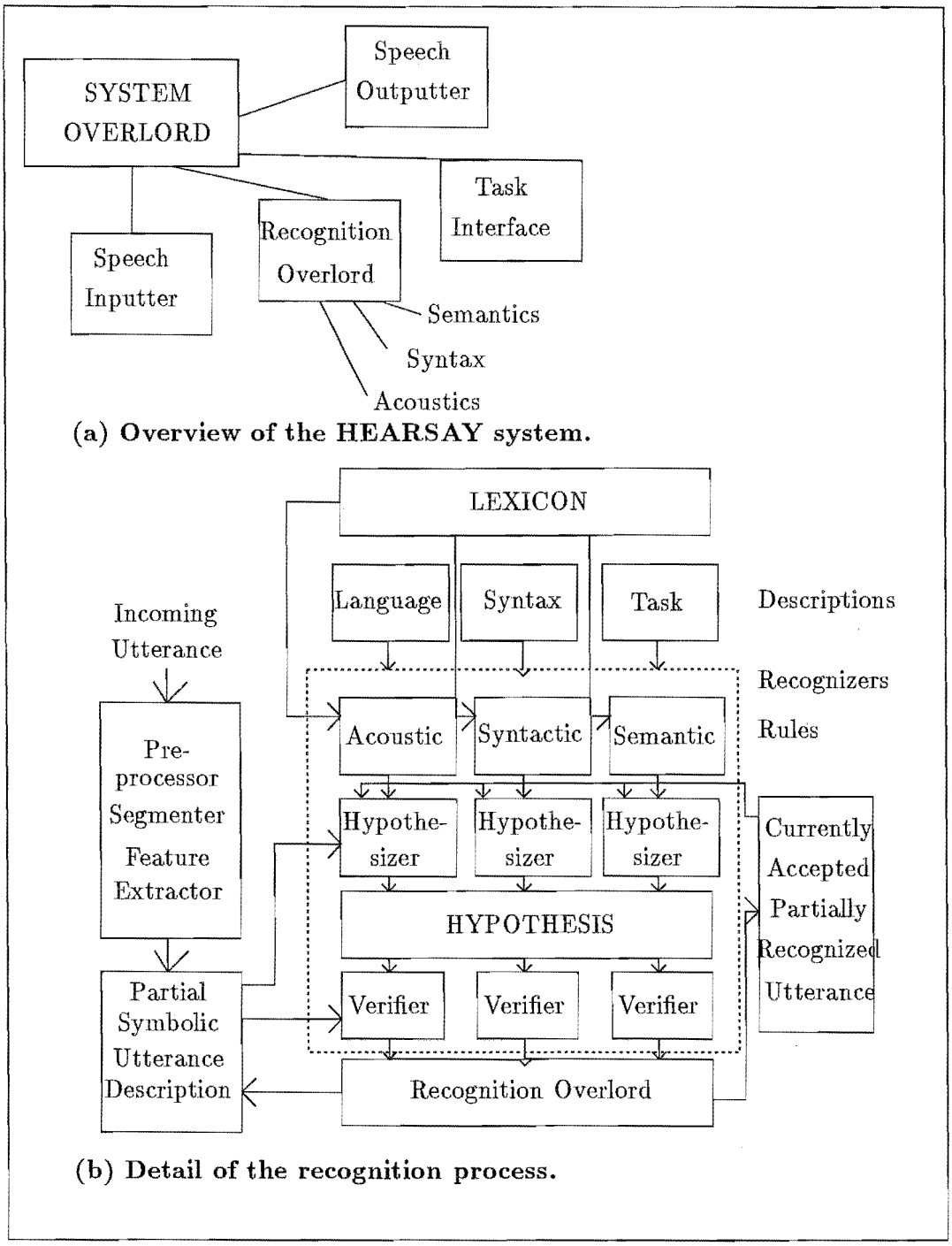


Figure 3.6. The HEARSAY system. (a) An overview of the HEARSAY system. (b) Details of the HEARSAY recognition process. (Reddy,1973)

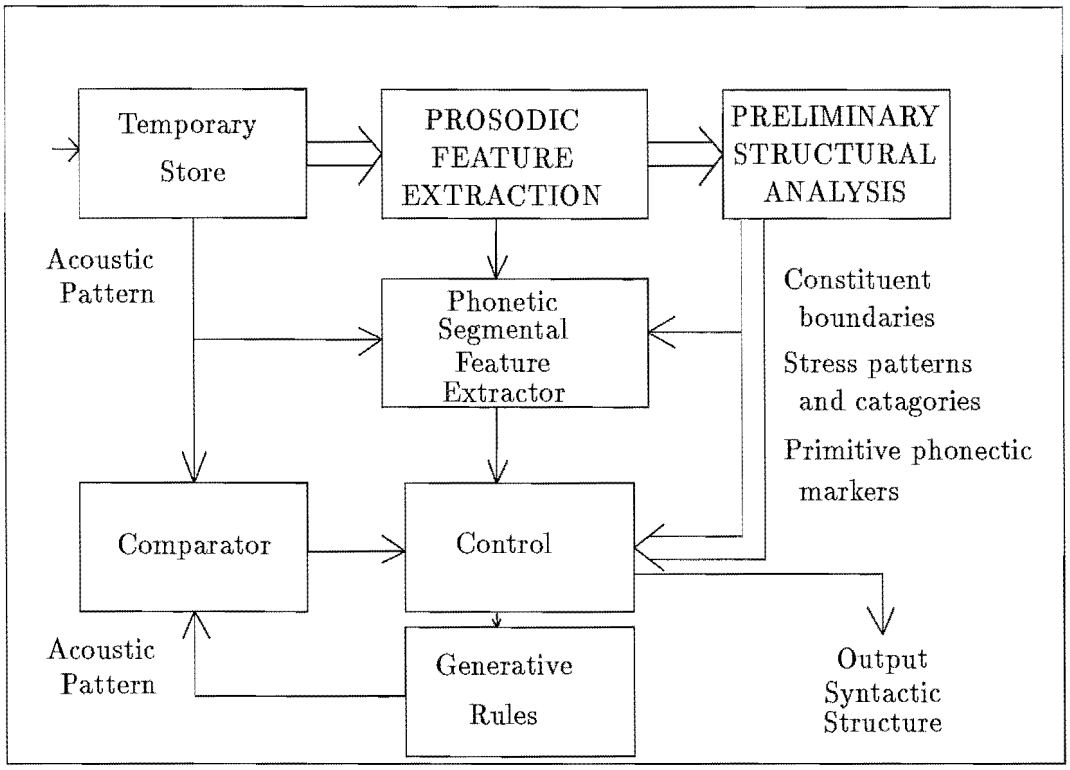


Figure 3.7. Application of prosodic features in an analysis-by-synthesis recognition scheme for syntactic recognition. (Lea,1973)

investigated the importance of prosody, such as pitch, on the understanding of speech sentences. By testing human recognition from spectrograms using prosodic information they found that the subject working with prosody performed significantly better than the control subject working without prosody. Lindblom and Svensson(1973) believed that by using grammar and prosody an unambiguous identification could be made in spite of an incomplete specification of the acoustic properties of the phonetic segments.

A more novel approach was taken by Baker(1974). Baker's Dragon system was a speech understanding system which used probabilistic information in the form of a Markovian process (refer §5.2). Using a Markovian process which assumes an external sequence dependent probabilistically on a internal unknown sequence, allows knowledge sources to be represented in a generative form. That is given the sequence of syntactic-semantic states one can generate the words, given the words one can generate the phones, and given the phones one can generate the acoustic information. The author noted that 'testing of the system is still at too preliminary a stage to make any definitive conclusions, but initial results are very promising'. This method of speech recognition was further developed in the 1980s (discussed in §3.3).

3.2.3 Summary

By the late 1970s much work had been undertaken into discrete and continuous speech recognition. Although discrete recognition was now giving useful performances, continuous recognition was still far from useful. To illustrate the problems with continuous speech recognition it is interesting to read an overview on the subject of recognition published by the military (Beek *et al.*, 1977). In this document, discussing mainly continuous speech recognition, the author emphasises the point that these systems are not

reliable. He also states the many problems with continuous speech recognition requiring 'drastically' restricted lexicon, syntax and semantics and usually only useful for a single speaker. The paper notes that something like 90% first choice accuracy is required to obtain good recognition for these systems and that the 'state-of-the-art is certainly two to three years from that goal'. The authors conclude by stating how a greater understanding of the speech process is required before automatic speech recognition (ASR) can approach human performance and that 'significant advances in ASR are not likely to come solely from research in pattern recognition, computer development and waveform/signal processing. Although these areas of investigation are important tools for ASR, the significant advances will come from studies in acoustic-phonetics, speech perception, linguistics and psychoacoustic equipment'. From the systems discussed, however, even though many systems were incorporating the large databases required for linguistic information the improvements in recognition accuracies were minimal. Many recognition researchers were again examining the acoustical information hoping to improve this area of accuracy.

3.3 THE 1980S

..the major problem in the field of speech recognition research has been, and still is, the difficulty of construction of an adequately complex, complete, and useable hypothesis to account for the extraordinary human ability to generate and interpret (understand) meaningful speech. (Hill, 1972)

The movement towards the construction of a useful recognition system grew substantially during the 1980s. However the ability to emulate the human was still far from adequate. Even though the research showed slow progress the growth of the word and speech recognition field in the 1980s was phenomenal. During the 1980s over 800 papers were published in the speech recognition field alone, almost double that published in the 1970s. A further indication clearly showing the growth of research in speech analysis and recognition is the expansion of the conference of the ASSP known as the *International Conference on Acoustic Speech and Signal Processing* (ICASSP). The first conference, in 1977, had approximately 220 presentations of which some 60 were speech related and 30 dealt with the problems of word and speech recognition. By 1989 the number of papers had increased so that over 700 papers were presented in which approximately 160 discussed speech problems and some 80 were directly related to speech and word recognition. From the 13 meetings between 1977 and 1989 over 2000 papers have been published on areas relating to speech processing(Mariani, 1989). What had begun, in the 1940s and 1950s, with some 10 laboratories world-wide dealing with speech and speech recognition had, by the late 1980s, grown to approximately 380 laboratories world-wide.

Much of the progress in the 1980s occurred within the fundamental areas of recognition systems referred to as *axes of difficulty* (Mariani, 1989). These axes of difficulty reflected the movement of recognition system from simple isolated word speaker-dependent systems to complicated speech recognition speaker-independent systems. The three axes along which the systems were moving were, firstly, the type of utterance recognised, moving from discrete word recognition to continuous speech recognition. Secondly, the speaker range from which the system could accurately recognise utterances, that is more systems were becoming speaker-independent, and thirdly, the vocabulary recognised with systems moving away from restricted vocabulary sets which limited the vocabulary to a few words and instead to recognise unlimited vocabularies.

Improvements in these three fundamental areas not only increased a recogniser's ability but also increased its complexity. More complicated systems were able to be

researched due to increased funding, the provision of standard databases of isolated and continuous words, and the expansion of computer technology.

The examination of these axes of complexity involved the investigation of many aspects of the recognition process including areas such as the methods of recognition (DTW, HMM or neural nets), the features chosen to represent the words, the methods of training and testing (clustering data, recognition decisions), the type of speakers (speaker-dependent, speaker-independent, speaker accent), the size and type of vocabularies(words, digits, letters), the type of speech (continuous, connected or isolated) and other associated parameters (noise, distance measures, coding). The growth of these areas during the 1980s will be discussed in the following sections.

3.3.1 The Methods of Recognition

Three major methods researched at this time were proving successful for recognition; dynamic time warping (DTW), hidden Markov modelling (HMM), and neural networks. The following section discusses these methods and the accuracies achieved in systems of the 1980s. Further details of the systems are given in Table A.3 in Appendix A.

3.3.1.1 Dynamic Time Warping

Although the DTW method was still the most widely used recognition method during the later part of the decade many HMM (and to a lesser extent neural network) recognition systems were being implemented. The DTW method was used predominantly as the recognition system for the testing of other variables such as acoustic features and distance measures (Rosenber *et al.*, 1983; Nocerino *et al.*, 1985; Morii *et al.*, 1985). A problem of the DTW method is its lack of speed due to the algorithm requiring a large amount of computation. This problem limited the use of DTW to simpler systems with small vocabularies. Many variations on the basic DTW method were tested to either increase recognition accuracy, widen the uses of DTW (to continuous recognition), or speed up the time consuming DTW process. Often special purpose hardware, in the form of VLSI circuits, was introduced to increase the speed of recognition schemes (Condick and Chalmers, 1989).

Methods to increase the accuracy of isolated word recognition DTW systems used schemes such as the *two pass* DTW method (Rabiner and Levinson, 1981; Tribolet *et al.*, 1982). The two-pass method operated such that on the first DTW pass a subset of the vocabulary is chosen, then, on the second pass, another DTW scheme is used to weight regions of the pattern increasing discrimination between words such as ONE and NINE (Rabiner and Levinson, 1981). This method, also used by Tribolet(1982), was incorporated into an isolated word LPC-based DTW recogniser. The discrimination weightings were calculated during the training phase of the recogniser and used in the second pass to increase recognition accuracy by 1 to 5%. Moore(1983) also used this method to distinguish words that sounded alike (such as STALAGMITE-STALAGTITE, FIVE-NINE) calculating the weightings based on distance score during the DTW algorithm. Recognition accuracy increased by 7% (26.8-19.8%). Other DTW weighting methods were used which weighted the transitions greater than the stationary parts of the speech and produced greater accuracies (Lamel and Zue, 1982; Watari *et al.*, 1983).

To speed up recognition, modifications to the standard DTW scheme were incorporated. These modifications were in the form of efficient search techniques, calculation reductions, or template reduction techniques. Search techniques such as the branch-and-bound, beam-search (Bisiani and Waibel, 1982), ordered graph search (Colla *et*

al., 1985; Brown and Rabiner, 1982) or metric-space search (also known as the approximating and eliminating search algorithm) (Vidal and Lloret, 1988) were introduced. Bisiani and Waibel(1982), Colla(1985), Brown and Rabiner(1982) and Vidal(1988) all used methods which searched for the best recognition template more efficiently. Using these methods recognition time could be reduced up to 50% without affecting accuracy.

During the 1980s the uses of DTW expanded from simple word recognition to that of connected and continuous recognition. For connected word recognition the level-building method (refer §5.1.5.1) was proving successful on unlimited vocabularies with 5-10% word errors for speaker-dependent recognition (Rabiner *et al.*, 1982). When tested for speaker-independent recognition of connected digit sentences (Rabiner *et al.*, 1986) the level-building method obtained 98-99% accuracy. Another connected word recognition technique developed during the 1980s was the *one-pass* method introduced by Bridle(1982). This method was so named because it boasted recognition and segmentation of the words during the same single pass. Another continuous recognition system is that of Yato *et al*(1986) which used DTW to recognise segmented phonemes from continuous digit sentences with 90% accuracy. Gauvain(1986) also used DTW for comparing sub-word units, in this case syllables, in his large scale word recogniser (recognising 10400 words). Word accuracies of 94% were quoted for speaker-dependent recognition.

Comparisons of DTW and HMM methods began appearing in 1983 when Levinson(1983) first compared the performance of a DTW system and a HMM system. Levinson(1983) found that both systems gave 3.5% error but that the errors occurred with different words. Rabiner(1983) also compared HMM and DTW recognition finding that HMM had slightly lower accuracy than DTW. Rabiner(1983) attributed the lower accuracy to the number of reference templates used to train the HMM system stating that only 100 reference utterances per word were used. Word recognition using HMM and DTW was again compared by Svendsen *et al*(1989) who obtained higher accuracy with the DTW method (97%) than the HMM method (91%).

3.3.1.2 Hidden Markov Modelling

HMM gained popularity in the later part of the 1980s because of its accuracy and computational speed. Rabiner first tested a HMM method in 1984 obtaining error rates of only 2%. Russell and Moore(1985) tested two different methods of HMM and also compared these methods with DTW. Both methods gave high accuracies however highest accuracies occurred with DTW for, what Russell and Moore(1985) stated, were 'easier' vocabularies while HMM methods performed better on word-pair confusion sets consisting of words that sounded similar. Juang *et al*(1985) tested two types of HMM methods as well a DTW method. They found that, on average, all three methods obtained equally high (>95%) accuracies. Tests with continuous recognition began with Mergel and Ney(1985) who tested a HMM method used for recognising phonemes in a continuous speech recognition system. Speaker-dependent and speaker-independent tests gave error rates of 4 and 13% respectively. Further testing undertaken by Rabiner(1989) for a connected digit recognition system using a standard Texas-Instruments (TI) connected digits database consisting of 225 adult speakers produced recognition errors of less than 2%.

HMMs were popular with large vocabulary systems, such as the SPHINX, the DECIPHER, the Lincoln continuous system and the IBM speech recognition system. The SPHINX system, discussed by Lee *et al*(1989) was developed at Carnegie-Mellon University. Results from tests on 150 sentences from 15 speakers gave accuracies as high as 82% without grammatical rules, and accuracies as high as 96% with word-pair grammar. Another large vocabulary system, known as DECIPHER (from Stanford Research

Institute) also used HMMs. In tests with 150 sentences from a standard database (from the National Institute of Standards and Technology(NIST)) accuracies of around 75% were achieved without grammatical rules, and 94% with grammatical rules. The Lincoln continuous speech recogniser was tested for both speaker-dependent and speaker-independent recognition on a 991 word task obtaining 3.5% and 12.6% error rates respectively. The IBM system discussed by Bahl *et al*(1989) was designed to recognise office correspondence in the form of continuously read sentences from a natural corpus covering a 5000 word vocabulary. Tests were conducted using 50 sentences read by 10 male speakers. Error rates for word recognition were 4% while for continuous speech sentence error rates varied between 11-26.8%.

A combination of DTW and HMM systems was proposed by Nakagawa *et al*(1986). The method for sentence recognition used a one-pass DTW system along with a HMM recogniser. The combination firstly recognised likely syllable strings using a one-pass DTW and then the particular string of syllables with the highest probability was chosen based on the HMM recogniser decision. Accuracies of 90-96% were quoted when recognising vowels of 90 Japanese names spoken by 3 Japanese male speakers.

Other uses for HMM systems have been the recognition of prosodic patterns of speech (Ljolje and Fallside, 1987) by characterising fundamental frequency, timing and intensity and also the recognition of cerebral palsy speech (Hsu and J.R.Deller, 1989).

3.3.1.3 Neural Networks

A third recognition method that became popular in the 1980s was that of neural networks. Neural networks were being tested for the recognition of simple sounds, individual phonemes, syllables and short words to avoid timing problems of words and sentences. Trehern *et al*(1986) tested a neural network for recognising the digits ZERO through FOUR. After 1000 learning cycles 100% recognition was claimed on the utterances used to train the systems and 64% on other utterances spoken by the same male speaker. A neural network recogniser's accuracy was compared to the accuracy of a simple distance measure recognition approach by Kamm *et al*(1989) for the recognition of single vowel sounds. A selection of distance measures were tested such as weighted cepstral, LPC likelihood, perceptual measures, and a distance derived by Kamm *et al*(1989) they called an *elastic* measure. Accuracies for the elastic and the weighted cepstral distance measures were highest for the distance measure method and comparisons with a neural network recogniser gave similar results.

3.3.2 Features Used

Many features were tested during the 1980s creating a great hotch potch of results but with very little useful comparative studies. Although the majority of the features tested were much the same as those used in previous decades these tests were undertaken with different recognition schemes involving DTW, HMM, and neural networks. The features encompassed frequency measures such as mel-frequency, linear frequency, cepstrum, LPC, parcor, LPC cepstrum (Davis and Mermelstein, 1980), and band pass filtered energies (Das, 1982; Moore *et al.*, 1983; Sugawara *et al.*, 1985; Iizuka, 1985; Mergel and Ney, 1985). LPC representations were still very popular (Tribolet *et al.*, 1982; Rabiner and Levinson, 1981; Rabiner *et al.*, 1982; Rabiner *et al.*, 1983; Pan *et al.*, 1985a; Bush and Kopec, 1985a; Ganesan *et al.*, 1986) as well as combinations of LPCs with energy calculations (Brown and Rabiner, 1982). One test by Rabiner(1984) which combined energy and LPCs in a single recogniser was reported to give a small but consistent accuracy increase of 0.2-1.0% over recognition with LPCs alone. Other combinations

such as zero-crossing rate and energy (Elghonemy *et al.*, 1986) were also tested giving small (1-2%) accuracy increases over the individual feature.

LPCs were used with both isolated and connected word recognition systems (Rabiner *et al.*, 1986). Cepstral derived features from LPCs (refer §4.4) (Morii *et al.*, 1985; Colla *et al.*, 1985; Togawa *et al.*, 1986) were also being used in isolated and connected recognition obtaining higher accuracies than LPCs (Juang and Rabiner, 1986). Juang and Rabiner(1986) tested both LPC and LPC derived cepstral coefficients with their HMM based recogniser. LPC accuracies were approximately 2-3% lower than cepstral accuracy. Mel-frequency cepstral coefficients (Schwartz *et al.*, 1985; Jouvet *et al.*, 1986; Gauvain, 1986) were also being used and producing 1-2% higher accuracies than LPC derived cepstral coefficients.

Temporal features, such as duration between successive zero-crossings, duration between extrema, and amplitude information (Baudry and Dupeyrat, 1982) and dynamic (transitional) information based on the movement of the feature representation from one frame to the next was popularised in the 1980s. Often the transitional and instantaneous features were combined (Soong and Rosenberg, 1988; Furui, 1986; Furui, 1989) to give further accuracy increases.

Modelling the characteristics of the human ear was also popular and, although this method was not new, features extracted from these models were now being widely used in recognition systems (Jelinek and others, 1985; Averbuch *et al.*, 1986). Hermansky(1985,1986,1987) designed a method which extracted perceptual LPC features based on an auditory model, claiming higher accuracies with less information for both speaker-dependent and speaker-independent recognition than recognition with LPCs.

3.3.2.1 Comparing Features

Few researchers were undertaking comparative studies of the features being proposed and those that did tended to discuss only a subset of the features resulting in a confusing mix of results that eluded classification.

The research team of Dautrich *et al*(1983), tested the use of filter bank features for word recognition. They compared filter bank feature extraction with the LPC method. Filter bank methods consisted of critical band, 1/3 octave band spacing, and flat and non-flat filter bands. Highest accuracies were obtained from LPC representation and critical band filtering. Partalo and Sijecic(1989) compared sets of speech features, examining filter banks, LPCs and time domain features (energy and ZC), for an alphabet vocabulary. Highest accuracies were given for LPCs (87.4%), lowest accuracy for the time domain features (73.8%). Junqua and Wakita(1989) compared cepstral coefficients and perceptual coefficients (PLP) with cepstral coefficients (with a projection distance measure) giving the highest accuracies. Svendsen *et al*(1989) compared LPC (likelihood distance), cepstrum and liftered cepstrum features obtaining highest accuracy with liftered cepstrum. Hunt and Lefebvre(1989) compared several acoustic representation using noisy and noise free speech. The features tested were cepstral coefficients, with and without quefrency weighting, mel-cepstrum, auditory based spectrum, transitional cepstrum and transitional and instantaneous representations together. Highest accuracies were from a subtle combination of many of these representations.

3.3.3 Methods of Training and Testing

Often, the method of training a recognition schemes was unique to the recognition system. In 1980, however, Rabiner and Wilpon(1980) classified these methods of training under three headings; casual training, training by averaging, and statistical training (such as clustering). Testing these methods Rabiner and Wilpon(1980) claimed higher

recognition accuracies from the averaging and clustering methods over the casual training method.

By improving the training of a system by clustering, as discussed by Rabiner and Wilpon(1980), recognition accuracies improved particularly when recognising larger vocabularies and more speakers (Rabiner and Levinson, 1981; Rabiner *et al.*, 1983).

3.3.4 Speaker-dependent versus Speaker-independent

Up until the 1980s most recognition systems were speaker-dependent ones. One reason for this was due to the difficulty in obtaining the large amount of data required, in the form of many different speaker representations, to train and test speaker-independent recognition systems. By the 1980s, with the increase of computer power, the ability to store larger databases was possible and, with the introduction of standard databases (which could be purchased in the later part of the 1980s), speaker-independent recognition became easier to carry out. Speaker-independent recognition was widely accepted as more useful then speaker-dependent recognition particularly in most commercial applications. Hence, researchers were striving to produce highly accurate speaker-independent systems.

As discussed previously, the disadvantage of speaker-independent systems is that they need a large database of different speakers to correctly train. A large database of speakers was required to successfully encompass the speech variations of the speakers who may use the recogniser. To produce speaker-independent recognisers Rabiner and Wilpon(1981) and Rabiner *et al*(1984) trained their systems with a database consisting of 100 talkers (50 male and 50 female). Their recognisers were then tested on another ten speakers using the alphanumeric (AN) vocabulary producing accuracies in the mid 80%. Using large numbers of speakers with difficult vocabularies, such as the alphanumeric vocabulary, gave reliable results that could be successfully reproduced in non-laboratory environments with a wide variety of speakers(Rabiner *et al.*, 1984b).

Bush *et al*(1985,1986) and Bush and Kopec(1987) produced a highly accurate speaker-independent system which they trained and tested on a standard Texas Instrument's database. The database comprised of American speakers (taken from many different regions) and non-American speakers. By testing American speakers with a system trained on non-American speakers and vice versa they obtain speaker-independent accuracies which ranged between 90-99%.

Many speaker-independent recognition systems were producing highly accurate results, rivalling results published for speaker-dependent recognition systems. In 1983 Rabiner *et al* produced 96.5% accuracy for their speaker-independent digit recogniser. Then, in 1984, Rabiner *et al* produced a speaker-independent digit recogniser, recognising digits over a telephone line. This system gave 93% accuracy. In 1985 Iizuka discussed his 15 digit plus word speaker-independent systems which produced accuracies between 96-97%. Also in 1985 Morii *et al* discussed a speaker-independent speech recogniser which recognised phonemes of speech. This system recognised phonemes with an accuracy of 81.4%. In 1986 Uikita *et al* published a connected digit speaker-independent recognition systems producing a string accuracy of 93.4% and word accuracy of 98.4%.

3.3.5 Vocabularies of the Eighties

3.3.5.1 Limited vocabularies

Although the research of speech and word recognisers has strived for unlimited vocabulary in reality the vocabulary of most recognisers is limited to increase the recogniser's accuracy. A recogniser's vocabulary can usually be limited due to the task it is required

to perform. In the 1980s recognisers' applications were often governed by commercial requirements and hence many recognisers recognised commercially viable vocabularies such as the digits, the alphabet, airline words or computer terms.

Systems that recognised digits were presented throughout the 1980s obtaining accuracies of around 95 to 98% (Shore and Burton, 1982; Rabiner *et al.*, 1983; Pan *et al.*, 1985a) and accuracies from 80% to 90% when recognised with the alphabet (the AlphaNumeric (AN) vocabulary) (Bisiani and Waibel, 1982; Rabiner and Levinson, 1981; Dautrich *et al.*, 1983; Nocerino *et al.*, 1985; Tribolet *et al.*, 1982; Das, 1982; Pan *et al.*, 1985b). Another commercially viable vocabulary was the airline word vocabulary(Rabiner *et al.*, 1982) producing error rates of approximately 12%, (Rabiner, 1984). A 54-word computer terminology vocabulary (Myers *et al.*, 1981) obtained 99% accuracy with speaker-dependent recognition and 96% with speaker-independent recognition.

By limiting vocabulary or choosing a simple vocabulary of distinguishable words, recognition systems could increase their recognition accuracy. The relationship between the vocabulary and the recognition accuracy was first discussed by Rabiner(1982). Rabiner(1982) showed how difficult it is to compare two different recognition systems tested on different vocabularies. Examining recognition accuracies from the literature showed that vocabularies such as the 10 digits, the 54 computer word, the 91 American states, and the 561 word and phrase vocabularies all obtain approximately the same accuracies and hence have the same level of recognition difficulty. In 1985 Russell and Moore(1985) also showed how accuracies change with changing vocabulary by testing their system on 3 unique vocabularies; the isolated digits, the alphabet, and a set of confusable word pairs. Testing two types of recognition systems, HMM and DTW, the authors obtained error rates for the digit vocabulary of 0.5% for DTW and 2.7% for HMM, while for the confusable vocabulary the error rates were 21.0% for DTW and 14.7% for HMM.

3.3.5.2 Large Vocabularies

Although unlimited vocabularies word recognisers were technically difficult to produce many large vocabulary recognisers were becoming so large as to appear to the user to be able to recognize an unlimited vocabulary. Recognition of very large vocabularies began in the mid 1980s with systems such as that by Hatazaki *et al*(1986) who recognised up to 5000 words. Accuracy was increased for such a system by using knowledge of the lexicon, and the probabilities of particular words following and preceeding other words. The system was evaluated on 200 phrases spoken by 4 speakers with an accuracy of 74-85%. Another large vocabulary word recognition system was discussed by Averbuch *et al*(1986). They designed a real-time 5000 word isolated recognition system which obtained an accuracy of 94%. The design of a future system which would recognise 20000 words was also discussed as possible.

3.3.5.3 Sub-words

Another way of increasing vocabulary size was to recognise sub-word units such as syllables (Lamel and Zue, 1982; Mercier *et al.*, 1982; Togawa *et al.*, 1986), phonemes (Gauvain, 1986; Yato *et al.*, 1986), diphones (Colla *et al.*, 1985), and demisyllables (Ruske, 1982; Rosenber *et al.*, 1983; Sauter, 1985). By recognizing a limited set of sub-word units an unlimited set of words could be recognized by combining the sub-words. Systems which recognised subwords produced accuracies between 60-80%. Comparisons of different sub-word units were undertaken by many researchers. One comparison was that of Rosenberg *et al*(1983) who examined word recognition using demisyllable and

whole word representations. The authors claimed 18-33% error with demisyllables and 6-15% with words. Morii *et al*(1985) comparing phonemes, vowels, semi-vowels and consonants obtained an average accuracy of 81.4% for phonemes, 90.6% for vowels, 78.0% for semi-vowels, and 71.9% for consonants. Jouvet *et al*(1986) tested recognition accuracies using a number of different types of sub-word representations, comparing recognition of words, phones and diphones. Testing with the digit vocabulary gave the highest accuracy with diphones (93%) while word and phoneme recognition accuracy was around 90%. Nakagawa(1986) tested a recognition system on different types of Japanese monosyllables; vowels, consonants and syllables. Recognition with 3 Japanese male speakers, speaking 90 Japanese names gave average recognition accuracies for vowels of 90-96% while consonants produced an accuracy of 50-63% and syllable accuracy was 48-62%. Although obtaining lower accuracies with subword systems the gains that could be achieved (such as increased vocabulary size with low memory overheads) pressured research into this area. Phoneme and other sub-word systems were also the first choice for continuous recognition (discussed in §3.3.6) however many further advances were needed before theses systems could be used fully.

3.3.6 Continuous/connected Recognition

Many continuous and connected word recognizers used the techniques of HMM and DTW while others centered on linguistical knowledge base information. The connected word recognisers were popular in the 1980s (due to their usefulness in many commercial fields such as telephone usage, data processing and booking procedures) however this usefulness was limited due to their restricted nature (requiring the speaker to pause between words). By the mid to late 1980s many more continuous recognisers were being seriously examined as these schemes became greatly demanded. At the top of the range of continuous (and connected) recognition were the HMM methods which were becoming particularly popular for continuous and connected recognition. The systems using HMM included the SPHINX system the DECIPHER system, the Lincoln continuous system, and the IBM speech recognition system (discussed in §3.3.1). These systems produced recognition accuracies up to 95% for small vocabulary speaker-independent systems and 70% for large vocabulary speaker-independent systems (Colla *et al.*, 1985; Bush and Kopec, 1985b; Rabiner *et al.*, 1986; Gagnoulet *et al.*, 1991).

Phoneme and other sub-word methods of recognition were becoming the center of most continuous recognition systems. These systems recognised sub-words and then used complicated grammatical and probability rules (such as uni-, bi-, and tri-gram probabilities) to form words and sentences. These highly complicated systems produced accurate recognition; however their usefulness was limited due to their large computer requirements. One method used by Ostendorf and Roukos(1989) used a stochastic model which reduced continuous speech to phonetics. Their method achieved greater than 75% recognition accuracy of the individual phonemes. Chow and Roukos(1989) from BBN Corporation reported accuracies of around 90% when using grammar based rules along with the 1000 word Darpa vocabulary. The KEAL system, a continuous speaker-independent recogniser, which recognised individual phonemes using acoustic analysis was discussed by Mercier *et al*(1989). Sentence recognition was achieved by using a context-free grammar sentence recogniser to obtain a sentence recognition accuracy of 81.5% for a test of 324 sentences, with 91% word recognition accuracy obtained for the 1287 words tested. Another continuous recogniser was that of Paeseler and Ney(1989) called SPICOS. This system used a stochastic language model based on trigrams, bigrams and unigrams of words achieving 91% accuracy on a 200 sentence test vocabulary. Real-time systems were being proposed late in the 1980s. One such real-time system is that by Murveit and Mankoski(1989) from SRI who proposed a real-time

HMM based large vocabulary (3000 word) continuous recognition system. The method used grammatical information by incorporating bigram language models. Even while the system was still under development a 20 000 word system was being proposed.

3.3.7 Distance Measures

During the 1980s it was realised that the distance measure chosen can have a profound effect on the accuracy of a recognition system. Researches began examining various distance measures used in recognition schemes and large effects were noticed when distance measures were changed. The first example of distance measures affecting the accuracy was in 1974 when Itakura(1974) devised a perceptual distance measure based on LPC parameters known as the *likelihood distance measure* (refer §4.3). The distance measure proposed by Itakura(1974) was unique in that it was a probability measure directly applicable to only one type of feature set, namely LPCs. Further trials during the 1980s with probability distortion measures and Euclidean distortion measures, using sets of LPCs, showed the probabilistic measure to give higher accuracies. Moore(1983) trialed these measures obtaining 8% increase of accuracy with the probabilistic measure over the Euclidean measure.

Many other distance measures were also trialed. Nocerino(1985) tested 6 different recognition distance measures the Itakura Saito(IS), log likelihood ratio(LLR), likelihood ratio(LR), Euclidean using cepstral coefficients (CEP), weighted likelihood ratio(WLR) and weighted slope metrics (WSM). The recognition errors for the LLR, WSM, LR, CEP distances were all approximately 8.5%, while the WLR measure gave 9% error and the IS measure gave 11.5% error.

Mansour and Juang(1988) tested a range of distance measures that formed a set known as *projection* distance measures. They claimed that these measures optimise recognition by reducing the distortion between non-noisy and noisy speech data. Mansour and Juang(1988) tested a system trained in different noise environments using both Euclidean and projection distance measures, and obtained higher accuracies with the projection (79%) than the Euclidean (42%).

3.3.8 Summary

Although most experimentation in the 1980s (much like the 1970s) produced large quantities of results, the lack of organisation and carefully planned experimentation produced results which lack statistical rigour and were largely unrepeatable. Any major attainments in the 1980s came about from careful testing of the many recognition parameters using standard techniques and standard databases which gave opportunities for repeatable experimentation. These standards allowed for the successful testing of the many variable parameters in speech recognition such as the methods of recognition, the vocabulary, the methods of training and testing, the effects of speaker and speech types, the effect of noise and other associated effects. Further to this, the comprehensive examination of features also allowed the improvement of recognition systems' accuracies in cases where the comprehensive testing was supported by standard databases and standard recognition procedures.

Along with the variables mentioned in this section there were many others which affected the accuracies of recognition systems. These other variables, such as noise cancelling, phonetic modelling, linguistical probabilities *etc.* were also being widely tested to improve accuracies, speed, and the vocabulary range of recognition systems.

The discussion in the previous Chapter has emphasised the large number of recognition schemes that have been produced over the previous decades. The vast number of schemes produced has been in direct relationship to the vast number of parameters

and environmental factors that effect the recognition outcome. Each parameter and factor has spawn many different recognition schemes thus producing, over the decades, thousands of different (successful and unsuccessful) systems. It is also due to this large number of recognition parameters and environmental factors that the testing of recognition systems is difficult and results unreliable. This historical overview has emphasised the wide number of systems producing stated recognition accuracies that are difficult to compare. Thus, when comparing results, it is important to take into account as many environmental factors and recognition parameters as possible, as all these may affect the outcome. And, when obtaining results, the testing procedures must be repeated so that statistical analysis can be performed to obtain accuracy mean and standard deviations.

Chapter 4

RECOGNITION FEATURES

Many types of feature have been employed to parameterise the speech signal. These features range from the easy to calculate temporal features such as energy and zero-crossings to complicated and time consuming frequency analysis techniques such as formant tracking and pole position tracking. The algorithms discussed in this thesis are constrained by the requirement of real-time operation. To fulfil this requirement only those features are examined that can satisfy the real-time conditions currently attainable by today's DSPs. This chapter discusses the calculation of speech features deemed appropriate in the sense implied above. These features are the subject of experimental investigations reported in Chapter 8 to ascertain their attributes in the context of speech recognition. The temporal features discussed are energy, (§4.1), and zero-crossing rate, (§4.2). Frequency based representation is in the form of linear prediction coefficients, (§4.3), cepstral coefficients, (§4.4), and transitional cepstral features,(§4.5). Perceptual linear prediction coefficients are discussed in (§4.6). The well known FFT frequency analysis has not been used as a feature. Although useful features can be extracted from the FFT and the FFT routine is both a well known and well used method the computation required to extract the useful features from the FFT data (such as pole tracking, formant tracking, or formant bandwidth estimates) has meant that the FFT routine is excluded. Standard frequency analysis as a word recognition tool is well covered in Ichikawa *et al.*, 1973, White and Neely, 1976, Dautrich *etal*, 1983, Partalo and Sijercic, 1988, and Hunt and Lefebvre, 1989.

4.1 ENERGY

For speech sounds it is desirable to calculate the time varying energy which represents the intensity variations during different speech sounds,

$$E(n) = \sum_{m=0}^{M-1} (w(m)x(n-m))^2, \tag{4.1}$$

where $w(m)$ is a weighting window of length M which isolates a segment of speech. The function $E(n)$ characterises the time varying amplitude properties of the speech signal. The type of window, $w(m)$, affects the outcome $E(n)$ because it acts as a filter on $x(n)$. To give equal weightings to all speech samples a rectangular window is used whereas for non-equal weightings a Hamming or other window is used. Fig. 4.1 illustrates the different energy signals that result from differently shaped windows.

The shape of the energy signal is also affected by the length of the window, M in (4.1). If M is small relative to the pitch period, the energy values will fluctuate rapidly. However if M is much larger than a pitch period, say 4 to 5 times larger, the energy will be very slow to change missing the speaker's vocal changes. In such cases the window act as a low-pass filter with its frequency cutoff depending on the length of the window. This effect is illustrated in Fig. 4.2.

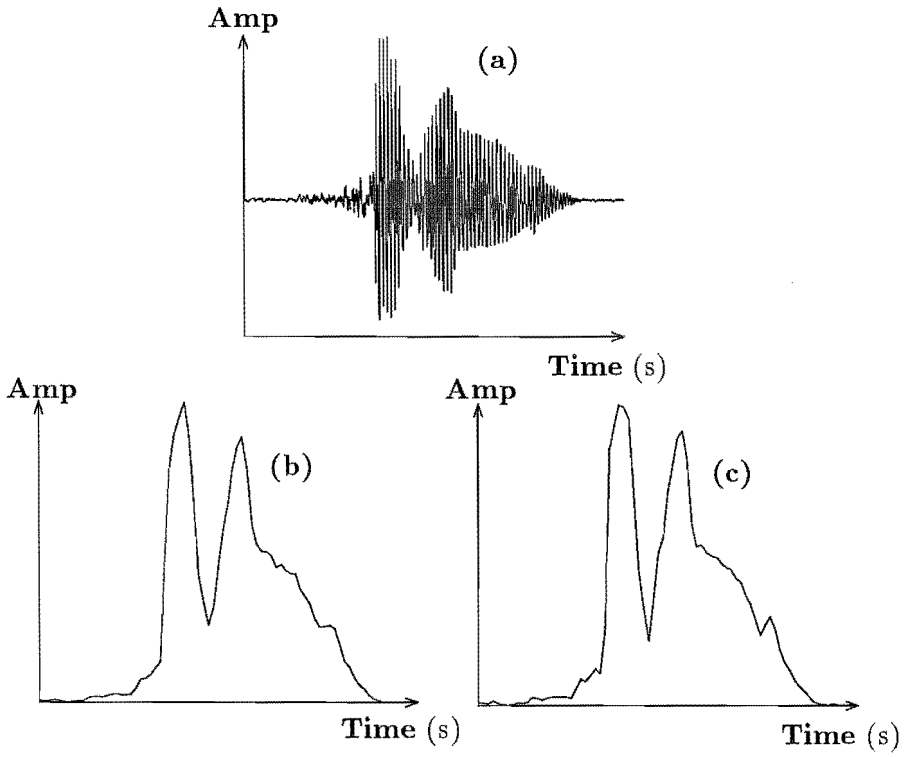


Figure 4.1. Plot of pressure waveform (a) and energy plots for the word SEVEN using two different window functions; (b)rectangular and (c)Hamming. Window length approximately two pitch periods.

One difficulty with the energy signal is that it is very sensitive to large signal levels because it is computed as the square of the sample values, thus requiring a large dynamic range. Often a root mean square calculation (RMS) is employed instead of energy as the preferred representation of speech loudness. RMS increases linearly with the magnitude of the speech samples, making it more useful for systems which have limited dynamic range, such as DSP fixed-point processors. The RMS profile of a digitised signal, $x(n)$ is defined as,

$$RMS(n) = \sqrt{\frac{\sum_{m=0}^{M-1}(w(m)x(n-m))^2}{M}}, \tag{4.2}$$

where $w(m)x(n-m)$ is the windowed speech sample and M is the number of samples in the frame or window. The advantage of the RMS is that it gives a simple, rapidly calculated feature that is capable of distinguishing words with unique envelopes, such as the words SIX and SEVEN. Unfortunately many other words tend to look very similar using this measure; examples of these are the words ONE and NINE. Both the speech and energy plots of the words SIX, SEVEN, ONE, and NINE are illustrated in Fig. 4.3.

A major function of the energy measurement is to separate speech sounds from non-speech sounds such as background noise, coughs and breath noise. Normally the energy of speech sounds is much greater than for non-speech noise. Rabiner and Sambur (1975) invoked this property to devise a detector for the beginning and ending points of words. From the energy contours of the words and by applying appropriate thresholds the beginnings and endings of the words were estimated. This method was also adopted by Lau and Chan(1985), who similarly used energy measurements to place

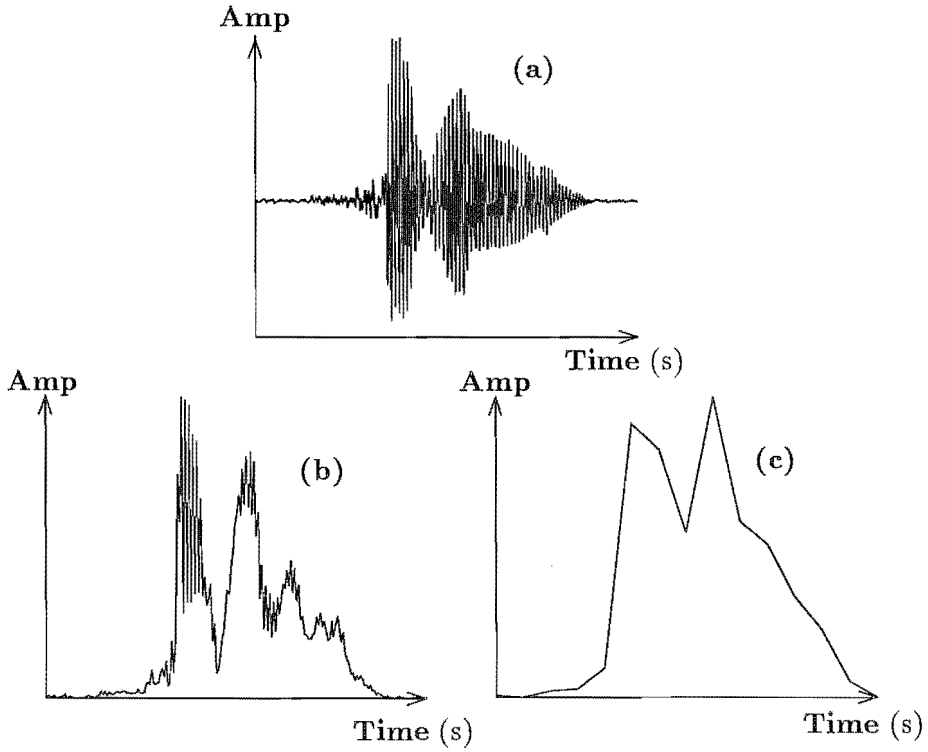


Figure 4.2. Plot of pressure waveform (a) and energy plots for the word SEVEN for two different window widths, (b) less than pitch period (c) greater than 5 pitch periods.

the beginning and ending points of words uttered in a sentence. Their method detected word endpoints when the energy of a speech frame dropped below five percent of the peak energy value across the word. Although energy has been successfully used for the detection of speech endpoints difficulties arise because of the effect of noise. Even low levels of noise, spikes from microphone knocks and breath noises from the speaker can cause errors in endpoint placement.

The energy signal has also been used for word recognition. When used as the only parameter for recognition reported recognition rates are in the region of only 30% (Rabiner *et al.*, 1984b). To improve accuracy many methods of combining energy measurements with other features extracted from the speech have been tried. Energy is often used as an addition to some frequency representation, such as LPCs, refer (§4.3),(Brown and Rabiner, 1982; Rabiner *et al.*, 1984b; Rabiner, 1984; Rabiner *et al.*, 1984a), or zero-crossings, refer (§4.2), (Lau and Chan, 1985). Brown and Rabiner(1982) combined energy and LPC information linearly in a frame-by-frame manner using the dynamic time warping method of time alignment, (the latter is fully explained in §5.1). The combination resulted in a 23.9% error reduction over LPCs alone. Rabiner *et al*(1984b) used energy and LPC in their isolated word recogniser in which separate energy and LPC distance metrics were used. The energy and LPC distances were added to obtain an overall value of the distance between test and reference words. A small but consistent increase of performance was found. Rabiner(1984) also proposed using energy with LPCs in a vector quantisation scheme, combining a weighted energy distance with the LPC Itakura distance. Word error rates dropped approximately 5% with this combination.

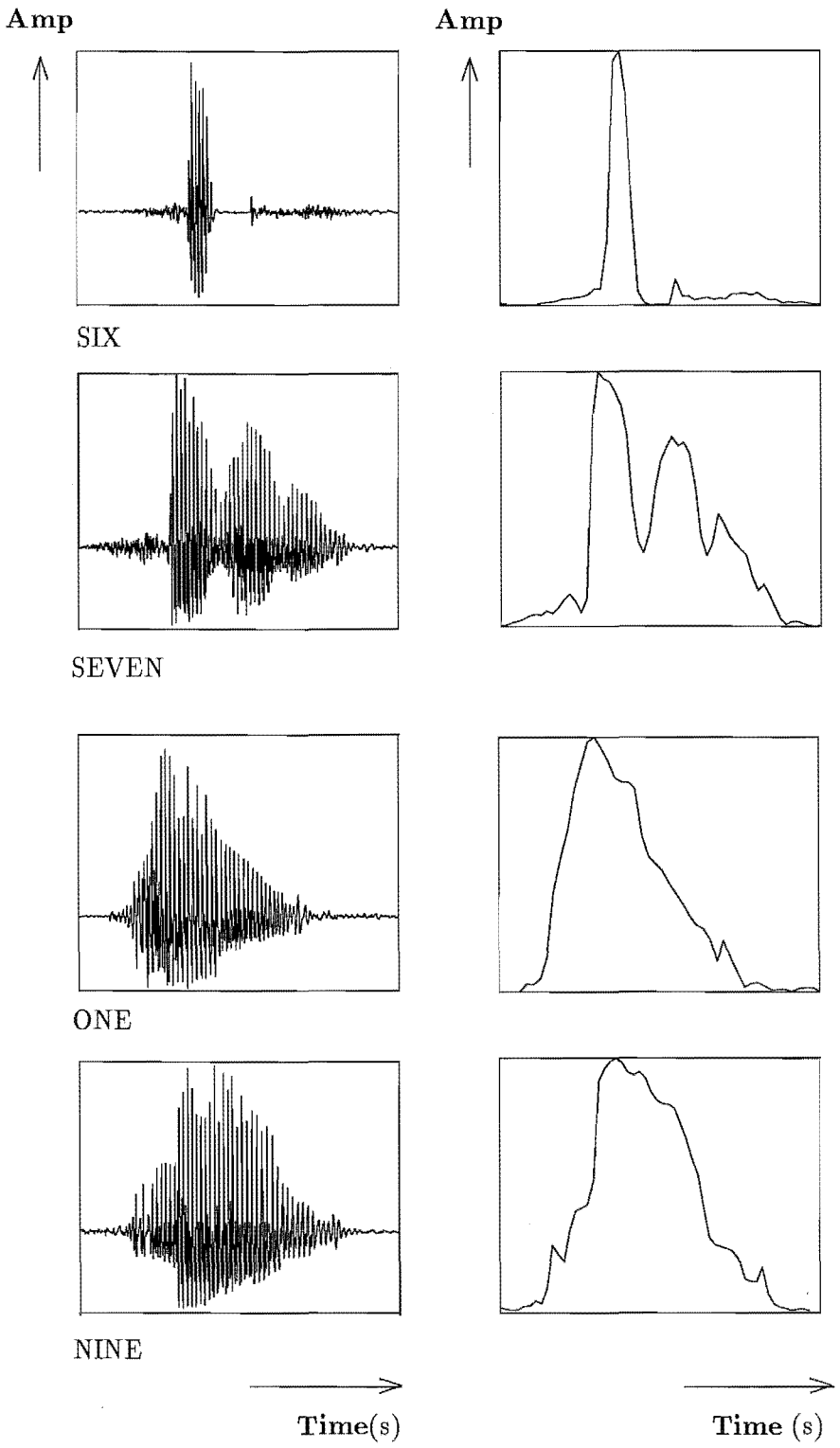


Figure 4.3. Speech pressure waveform and RMS plots of the words SIX, SEVEN, ONE, and NINE. The energy (RMS) plots for the words SIX and SEVEN are easily distinguishable while the energy (RMS) plots of the words ONE and NINE are very similar.

4.2 ZERO CROSSING RATE (ZX)

It was established by Licklider *et al.*(1948) that infinitely clipped speech, (that is speech that retains only its zero-crossing positions), remains highly intelligible to human listeners. Testing with clipped, differentiated clipped, and integrated clipped sounds showed that listeners were slightly better at recognizing the differentiated clipped speech than the clipped speech and up to nine times better at recognizing the differentiated clipped speech than the integrated clipped speech. It was soon realised that a high level of information was contained in the temporal pattern of the crossings of the time axis in the speech wave. In fact clipped speech is approximately 90% intelligible (Niederjohn *et al.*, 1987). Niederjohn *et al*(1987) also showed that there exists a strong proportional relationship between speech intelligibility and the percentage of zero-crossing locations preserved.

Zero-crossing measures have been widely used as features in speech recognition systems (Bezdel and Chandler, 1965; Lavington, 1969; Bezdel and Bridle, 1969; Ewing and Taylor, 1969; Niederjohn, 1975; Lau and Chan, 1985). The number of zero-crossings during some time interval (the zero-crossing rate) is the most widely used zero-crossing measure and represents the average zero-crossing information in speech (Niederjohn, 1975). The average zero-crossing rate, $ZX(n)$ can be written as,

$$ZX(n) = X(n)/M, \tag{4.3}$$

where $X(n)$ is the number of zero crossings in the frame n, and M is the number of samples in the frame. Several other methods of parameterising the zero-crossing information have also been implemented in recognition schemes (Niederjohn, 1975).

The time intervals between successive crossings of the zero line are related to the frequency components present in the sound, with the zero-crossing rate, $ZX(n)$ being strongly related to the dominant frequency component. It is, therefore, possible to obtain a frequency related measure by recording the rate of these crossings within a frame of the sound. For example, if the signal consists of a single sinusoid with frequency f_o, then the number of zero-crossings per second (ZX/s) is related to frequency by the relationship,

$$ZX/s = 2f_o. \tag{4.4}$$

The frequency interpretation of zero-crossing rate of actual speech is not so straightforward due to the interactions of the many sinusoidal components. Several researchers have attempted to relate the two measures (Ainsworth, 1967; Scarr, 1968; Niederjohn, 1975) but these relationships are not yet practical for speech recognition purposes.

One principal application of this measure is in separating the relatively low frequency voiced sounds from the high frequency unvoiced sounds (refer §2.1.2). Since high frequency sounds imply high zero-crossing rates and low frequency sounds imply low zero-crossing rates these sounds are usually separable. Fig. 4.4 gives a distribution of zero crossing rates (ZXR) for voiced, unvoiced and silence taken from the vocabulary ZERO through NINE. For this vocabulary the unvoiced distribution in Fig. 4.4(b) is relatively uniform, consisting of a reasonably large amount of low frequency sounds, which is unusual. The low frequency component in the unvoiced distribution is due to unvoiced sounds such as plosives (refer §2.1.2) found in the words TWO and EIGHT which have low frequency components.

Zero-crossings also have the ability to distinguish unvoiced sounds from background silence making this measure useful for endpointing words. However the zero-crossing rate can be highly affected by dc offset, any noise that may be present in the digitisation procedure, or any unwanted noises produced by the speaker such as breath noise and lip smacks. Separation of voiced and silence is difficult using ZXR because the distributions

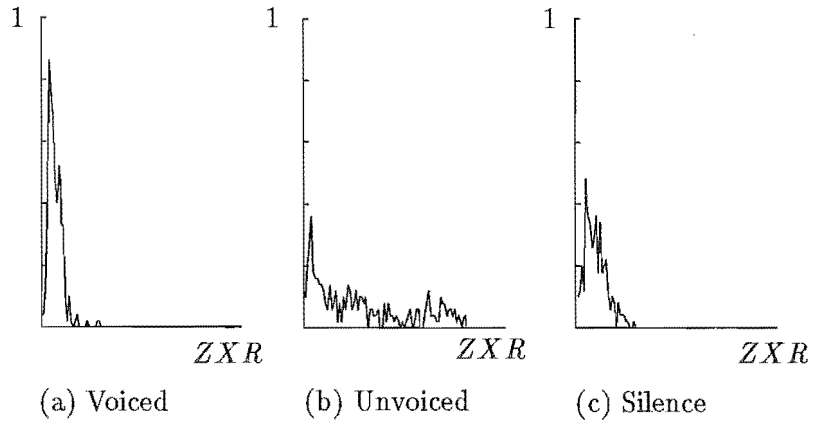


Figure 4.4. Probability density distributions of zero-crossing rates for (a) voiced sounds (b) unvoiced sounds (c) silence.

are similar. Identification of the endpoints is usually performed in conjunction with an energy measure to alleviate these problems (Rabiner and Sambur, 1975; Savoji, 1989), and in general this type of endpoint estimation gives (on average) 7-10% endpoint errors (endpointing errors are said to occur when endpoints are missed or misplaced with respect to a human recognizing the best place endpoints).

Zero-crossing information has been used as an effective feature for speech recognition. Bezdel and Chandler(1965) tested a zero-crossing method on vowel sounds and then Bezdel and Bridle(1969) tested this method on a vocabulary of 15 words. Their method recorded the distribution of the zero-crossing rate, also known as the probability density distribution (shown in Fig. 4.4), for each word. Recognition accuracies of around 96% for vowels and words were quoted.

Usually zero-crossing rate is not enough information for accurate recognition and in such cases it is combined with other recognition parameters. Lavington(1969) combined zero-crossings with rate of zero-gradient or 'turn-arounds' to achieve a recognition accuracy of 96%. Lau and Chan(1985) combined zero-crossing rate with energy (refer §4.1) to recognize ten Cantonese digits with an accuracy of 97.2%. Because no recognition accuracies were given for any of these recognition parameters on their own it is not known whether the combination of parameters improved accuracies.

4.3 LINEAR PREDICTIVE CODING (LPC)

LPC is one of the most common techniques used in the processing of speech and, in particular, in the field of speech recognition. The importance of this method lies in its ability to provide easily calculable and accurate estimates of the speech spectral envelope. This method enables the coding of large amounts of information about the speech with very little data, making it attractive for speech recognition.

The LPC coefficients are so named because a particular speech sample can be approximated as the sum of past speech samples each multiplied by a coefficient value. Hence a forward speech sample is *predicted* by a combination of past speech samples. The values of the coefficients are calculated by minimising the sum of the squared differences (over a finite interval) between the actual speech samples and the predicted ones. A unique set of predictor coefficients can be determined for each set of speech samples.

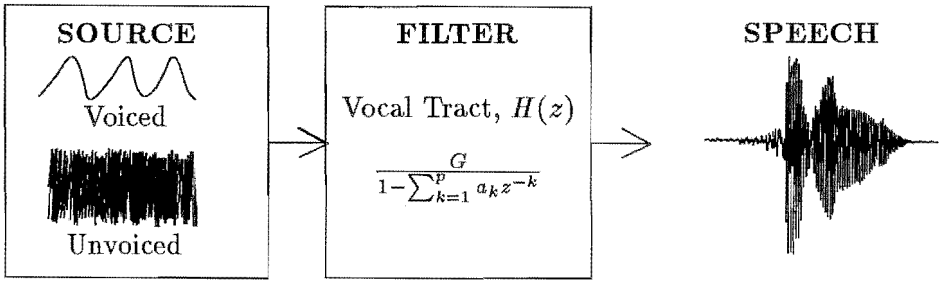


Figure 4.5. Basic speech production model showing source-filter representation. The source models the type of excitation. The filter models the vocal tract.

The LPC method represents the vocal tract as a linear, slowly time varying all-pole filter. In the z-domain the set of LPC coefficients, $\{a_k\}$, can be written as the filter coefficients of the vocal tract model $H(z)$. The vocal tract representation using a *pth* order all-pole filter model can be written as

$$H(z) = \frac{G}{1 - \sum_{k=1}^{p} a_k z^{-k}}, \tag{4.5}$$

where G is a factor which depends upon the area of the vocal tract and is defined in 2.3.3.

Describing the speech with LPC coefficients results in loss of information which must be replaced if the speech is to be reconstructed. The information which must be added consists of voiced/unvoiced classification, pitch period for voiced speech, and gain parameter G. These parameters, along with the $\{a_k\}$s are required for each frame in order to reproduce the speech signal. Fig. 4.5 shows the source-filter model which forms the basic model of the vocal tract and filter excitation. Voiced sounds are produced by exciting the filter model with a periodic pulse of energy, whose period is equal to the pitch of the sound. Unvoiced sounds are produced with a random noise excitation.

There are many methods to derive the predictor coefficients from the speech samples. The most useful methods are the autocorrelation method, the covariance method, and the maximum likelihood method (Fant, 1960; Markel and A.H. Gray, 1976; Rabiner and Schafer, 1978).

The autocorrelation and the covariance methods are both derived from the method of minimum variance. Assumptions of randomness and stationarity are made on the speech samples but no statistical function (such as Gaussian) is attributed to the speech. For these methods an error $e_n(m)$ can be defined with respect to the real signal $s_n(m)$ and the predicted signal $\sum_{i=1}^{p} a_i s_n(m-i)$, such that,

$$\begin{aligned} e_n(m) &= -s_n(m) + \textstyle\sum_{i=1}^{p} a_i s_n(m-i) \\ &= \textstyle\sum_{i=0}^{p} a_i s_n(m-i), \end{aligned} \tag{4.6}$$

for a pth order model with $a_0 = -1$ and where $s_n(m)$ denotes the mth sample in a speech segment n. The short time average prediction error, calculated over a sample set $\{m\}$ is defined as,

$$\begin{aligned} E_n &= \textstyle\sum_m \left(e_n(m)\right)^2, \\ &= \textstyle\sum_m \left(-s_n(m) + \sum_{i=1}^{p} a_i s_n(m-i)\right)^2. \end{aligned} \tag{4.7}$$

By minimising the squared error with respect to each of the predicted coefficient values the optimum set of coefficients can be calculated for the segment n.

The range of m, the number of samples used to predict the LPC values in (4.7), has not been stated and its selection leads to assumptions which must be made about finite sets of observed samples. Optimally m should range from $-\infty$ to $+\infty$ but in reality data of window lengths are used so this range is not possible. Setting m within the range $0 \leq m \leq N-1$ and assuming the waveform to be zero outside the interval leads to the autocorrelation method. Assuming the waveform is zero outside the boundaries of m can cause large prediction errors at the beginning and ending of the interval. At these positions the method is trying to predict non-zero speech values from samples of zero magnitude at the beginning of the waveform and zero values from non-zero values at the end of the waveform. To avoid this problem a window should be used which tapers the samples slowly to zero avoiding any sudden magnitude changes at the boundaries. The autocorrelation method has the advantage of always being solvable and stable, and efficient algorithms are used for calculation of the predictors for this method such as the Durbin-Levinson algorithm (Markel and A.H. Gray, 1976).

Setting m within the range 0 to $M-1+p$ (where M is the length of a frame of speech data) leads to the covariance method. To evaluate the covariance function requires sample values in the interval $-p \leq m \leq M-1$. Extending calculation to $-p$ means that initial predicted values, calculated from $m=0$, have previous sample values to calculate from. This means that this method does not have to use a window function because it is not assumed that the signal is zero outside the region of interest. However this computational method does not guarantee a stable system and in fact may have no true solutions.

The maximum likelihood method of LPC computation (first used by Itakura and Saito(1970)) begins by modelling the speech process by a Gaussian, stationary, random process that is generated by passing uncorrelated white Gaussian noise through an all-pole model. These assumptions may not seem completely valid for actual speech but the formulation proves to be equivalent to the other methods such as the autocorrelation method described previously(Itakura and Saito, 1970). A Gaussian process is assumed because the system is represented by a random process being shaped by a linear system (also known as the central limit theorem) (Markel and A.H. Gray, 1976). The process is also considered to be stationary in that its statistics are assumed not to vary over time - which is reasonable over short periods of unvoiced speech. However, these assumptions may not be valid for all types of sounds, voiced and unvoiced(Markel and A.H. Gray, 1976; Rabiner and Schafer, 1978).

It is also possible to calculate the reflection coefficients, as discussed in §2.3.2, from the linear prediction coefficients. The recursive solution used to efficiently calculate the autocorrelation based linear prediction coefficients produced a set of reflection coefficients as intermediate variables. Markel and Gray(1976) show that the reflection coefficients of the autocorrelation method are equivalent to the acoustic reflection coefficients of an acoustic tube model of the vocal tract when the speech is both sampled at an appropriate rate and pre-emphasised before analysis. It is beyond the scope of this thesis to derive this relationship however Markel and Gray(1976) Chapter 4 is an excellent reference for further information on how the reflection coefficients can be calculated and the recursive linear prediction algorithm.

Many of the assumptions of the vocal tract model are not valid under very basic conditions. One assumption is that the glottal excitation of one period does not interact with the following speech period. Markel and Grey(1976) quote from Atal(1974) to show that with increasing fundamental frequencies errors in formant estimation increase and that this is because the excitation has a much greater interaction with the next

period of the speech. Errors also increase in the formant bandwidth estimation with errors increasing as fundamental frequency increases. The result of this is explained by noting that for increased fundamental frequencies the major oscillation within a period will have interference from the decaying oscillation of the previous period with the greatest error occurring when the formant is multiples and a half of the fundamental (ie formant = $F_0/2$ plus multiples of F_0) with error also increasing as fundamental increases due to reinforcement from previous periods.

Another assumption is that the speech is formed from an all-pole model and hence can be represented by an all-pole system. Those sounds which contain spectral nulls, such as nasalised sounds are therefore badly represented. This is usually not considered a problem because poles in the spectrum are perceptually more important than zeros and should therefore be modelled with greater accuracy (Fant, 1960; Rabiner and Sambur, 1976). However, frequency nulls cause effects such as bandwidth widening of the formants which, if the effects are large enough, can be important and so should be accurately modelled. It has been noted that recognition rates for nasalised sounds can be greatly reduced if these properties are not sufficiently well modelled (Juang *et al.*, 1987). Juang *et al*(1987) noted LPC analysis on this type of sound gives results which vary significantly for nominally similar signals, particularly around the region of the spectral zeros. One way of determining the accuracy of the LPC analysis is to examine the spectral output from the analysis. The success of formant estimation for a particular number of speech data points depends strongly upon the particular speaker, utterance, and speaking conditions. Best results can be predicted for a male speaker with low fundamental frequency producing a loud vowel sound (Markel and A.H. Gray, 1976). Markel and Gray (1976) listed a number of conditions which lead to unreliable results such as high fundamental frequency, insufficient interval of glottal closure, and utterances of low intensity.

A limitation of LPC analysis is that the number of prediction coefficients must be chosen *apriori.* The actual number of prediction coefficients that are needed to accurately represent the speech is determined by a multitude of factors which change with the sounds produced, such as vocal tract length, nasal cavity coupling, and the excitation (Fallside and Woods, 1983). Because the number of coefficients must be chosen prior to calculations usually the number is chosen to be dependent on the delay which occurs when sound travels the length of the tube. The delay of the filter tube model determines the number of components because the components relate to a filter model (refer equation 4.5) and so the delay equates to the time the signal takes to propagate through the filter. Because the system represented by a filter model is a sampled (digital) system this can be related to the sampling frequency of the system. For a sampled system (which is usually the case) the delay can be equated to the inverse of the sample rate of the system. For a tube with delay T of $\frac{1}{2F_s}$ for sampling frequency F_s, taking speed of sound c and vocal tract length L, then the number of tube sections, which equates to the number of predictor coefficients, p, can be calculated with respect to the delay through the tube $(\frac{1}{2F_s})$ such that

$$p = \frac{2LF_s}{c}. \tag{4.8}$$

For an adult male with a vocal tract length L of 17cm and sampling at 10000 Hz, p =10. The fixed number of poles thus enforced can lead to either lack of representation of all the frequency components or else the addition of spurious resonant peaks due to too many poles being modelled (Kirkland, 1993) as shown in Fig. 4.6.

Modelling errors can also be associated with the frame length over which LPC coefficients are computed. This relationship between framesize and error is highly important for the autocorrelation method which uses a window to weight the speech

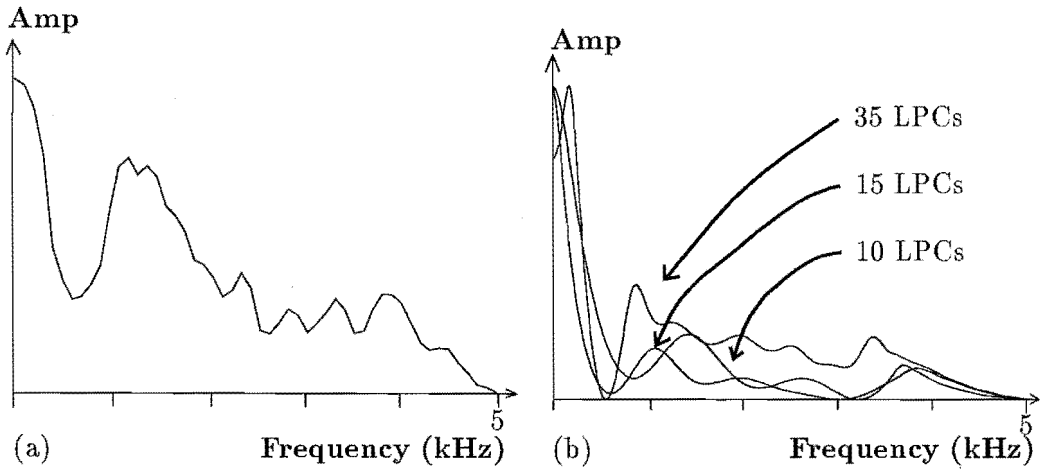


Figure 4.6. Comparison of frequency spectra calculated by (a) short-term FFT analysis (b)LPC analysis

samples. For this case the duration of the frame length must be long enough so that the tapering effects of the window do not unduly affect the results. It has been shown that the frame length M must be of the order of several pitch periods to ensure reliable results (Markel and A.H. Gray, 1976). Thus analysis durations of 100 to 400 samples (at 10kHz sampling rate) are required with most systems favouring the larger values of frame length (Chandra and Lin, 1974; Rabiner and Schafer, 1978).

Poles calculated from linear prediction analysis can be mapped into the frequency domain. Deriving the frequency envelope from the LPC coefficients allows the calculation of an equivalent short-time (less than one pitch period) spectrum which removes the fine structure contributed by the pitch frequency. A representation of the frequency of a vowel sound is shown in Fig. 4.6 using both a short-time Fourier analysis which removes the structure and LPC analysis. The LPC analysis is shown with several different numbers of LPC coefficients. As can be seen from Fig. 4.6 the greater the number of LPC coefficients employed to represent the speech the greater the number of peaks found in the spectrum and the smaller the bandwidth of each peak. Thus the LPC representation can produce more definitive formant structure. However with too many LPC coefficients extra formants may be added making the modelled formant structure unreliable.

The first researchers to directly apply linear prediction techniques to speech analysis were Itakura and Saito(1968) and Atal and Schroeder(1968). The use of LPC coefficients for speech recognition began in the early 1970s. By 1973 Ichikawa had compared LPC coefficients with other standard recognition techniques such as spectrum (derived from band-pass filtering), cepstrum (refer §4.4), autocorrelation coefficients derived from the time waveform and partial autocorrelation coefficients. For these tests LPC coefficients gave up to 30% lower recognition accuracy than the other features tested. In 1974 Itakura claimed a recognition accuracy with LPC coefficients of 97% for a set of words and 88.6% recognition with an alphanumeric vocabulary. There appeared little difference between the methodology of these two tests however the vocabulary and conditions in which the tests were undertaken were different and may have been the cause of the differing results. Following Itakura's tests LPCs gained popularity in isolated word recognition schemes producing accuracies around 85-95%(Sambur and Rabiner, 1976; Rabiner and Schafer, 1978; Rabiner and Schmidt, 1980; Rabiner *et al.*, 1982; Rabiner *et al.*, 1984b; Partalo and Sijercic, 1988). Recognition experiments by White

and Neely (1976) gave approximately the same results for both LPC coefficients and spectrum (produced from bandpass filtering) of around 95%. LPCs as recognition parameters for continuous and connected speech produce error rates of 2-3% (Rabiner and Schmidt, 1980), to 5% (Medress *et al.*, 1976). Used for the segmentation of sentences into words with other parameters such as energy (refer §4.1), zero-crossing (refer §4.2), and autocorrelation coefficients derived from the time domain waveform, a segmentation accuracy of 99% has been achieved (Rabiner and Sambur, 1976). Christiansen and Rushforth(1977) used LPCs to detect and locate specific words in continuous speech with an average accuracy of 97%.

4.4 CEPSTRAL COEFFICIENTS(CEP)

Cepstrums have been used widely in speech research as a tool to reduce any two signals which are convolved in the time domain into two signals which add in the cepstral domain. This method allows the separation, by subtraction in the cepstral domain, of two signals convolved in the time domain.

The complex cepstrum is the inverse Fourier transform of the complex logarithm of the Fourier transform of the input,

$$\begin{aligned} \hat{x}(n) &= \tfrac{1}{2\pi}\int_{-\pi}^{\pi} log[X(e^{jw})]e^{jwn}dw, \\ &= \tfrac{1}{2\pi}\int_{-\pi}^{\pi} \hat{X}(e^{jw})e^{jwn}dw. \end{aligned} \tag{4.9}$$

The cepstrum is a specific case of a signal which is defined by a homomorphic system (Gold and Rader, 1969). Conventional linear homomorphic systems obey the principle of superposition which states that if an input signal is composed of a linear combination of elementary signals, then the output is a linear combination of corresponding outputs such that, if L represents a linear operator,

$$\begin{aligned} L[x(n)] &= L[x_1(n) + x_2(n)], \\ &= L[x_1(n)] + L[x_2(n)], \\ &= y_1(n) + y_2(n), \\ &= y(n). \end{aligned} \tag{4.10}$$

However, in the case of cepstrums the signal combination is by convolution, represented by the symbol $\odot$ and, by analogy to the linear combination above, a generalized principle of superposition can be obtained where the addition is replace by convolution. An important aspect of the theory of homomorphic systems is that any homomorphic system can be represented as a cascade of three homomorphic systems. For the case of homomorphic systems for convolution the first homomorphic system takes inputs combined by convolution and transforms them into an additive combination of corresponding outputs. The second system is a conventional linear system obeying the principle of superposition as given in 4.10. The third system is the inverse of the first system (it transforms signals combined by addition back into signals combined by convolution). The importance of such homomorphic systems is that the design of such systems reduces to the problem of the design of linear systems. The operation which converts an input, which is convolution, to an output, which is ordinary addition, is known as the characteristic system for homomorphic deconvolution (Rabiner and Schafer, 1978) and obeys a generalized principle of superposition. This characteristic system for homomorphic deconvolution is defined as

$$C[S_1(t) \odot S_2(t)] = C[S_1(t)] + C[S_2(t)]. \tag{4.11}$$

To illustrate the cepstral process two signals are examined which, in the time domain, are convolved together,

$$x(n) = x_1(n) \odot x_2(n). \tag{4.12}$$

In the frequency domain (4.12) becomes the multiplication,

$$X(e^{jw}) = X_1(e^{jw}).X_2(e^{jw}), \tag{4.13}$$

where the upper-case letters correspond to the spectra of the lower-case time domain signals. The cepstral process now reduces the signal combination to addition by using logarithms,

$$\begin{aligned} \hat{X}(e^{jw}) &= log[X_1(e^{jw}).X_2(e^{jw})], \\ &= log[X_1(e^{jw})] + log[X_2(e^{jw})]. \end{aligned} \tag{4.14}$$

If the signal is complex the complex logarithm must be used so that,

$$\hat{X}(e^{jw}) = log|X(e^{jw})| + jarg[X(e^{jw})]. \tag{4.15}$$

Note that the complex logarithm function does not produce a unique output. In fact the complex logarithm violates (4.14) because the principal value of the logarithm of a product of complex signals is not always the sum of the principal values corresponding to the individual complex signals. Nevertheless restrictions, in the form of signal constraints, placed on the complex input signal can produce a satisfactory complex output and so the complex cepstrum can be found (Gold and Rader, 1969).

There is however, no ambiguity in the real part of a complex logarithm, it is always equal to the logarithm of the magnitude of the complex signal. The real cepstrum can then be found as the inverse transform of the real logarithm and is also defined as the inverse transform of the log spectrum of the magnitude of the Fourier transform, (for speech this is equivalent to the Fourier transform of the log of the power spectrum), and can be written as,

$$c(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} log|X(e^{jw})| e^{jwn} dw, \tag{4.16}$$

which is equal to the even part of the complex cepstrum $\hat{x}(n)$.

For speech the real cepstrum is all that is usually computed. Since the speech waveform is usually sampled before processing to form $x(n)$, the real, discrete cepstrum, $c_p(n)$ must therefore be calculated using a discrete formula. The formula uses N samples of the discrete signal in the time domain to produce K samples in the frequency domain which is transformed to N samples in the cepstral domain such that

$$\begin{aligned} X_p(k) &= \textstyle\sum_{n=0}^{N-1} x(n) e^{-j\frac{2\pi}{N}kn}, & 0 \le k \le K-1, \\ C_p(k) &= log[|X_p(k)|], & 0 \le k \le K-1, \\ c_p(n) &= \tfrac{1}{N} \textstyle\sum_{k=0}^{K-1} C_p(k) e^{j\frac{2\pi}{N}kn} & 0 \le n \le N-1 \end{aligned} \tag{4.17}$$

The resulting discrete cepstrum $c_p(n)$ is an aliased version of the true cepstrum $c(n)$ such that,

$$c_p(n) = \sum_{r=-\infty}^{\infty} c(n + rN). \tag{4.18}$$

This aliasing effect would appear to make it necessary to use a large value of N (generally a large value is greater than 512) to reduce the effect of aliased overlapping of discrete cepstrum. Usually, however, large values of N are not selected because of limitations of the amount of data available and the requirement to use short-term information for speech recognition procedures. Thus smaller values of N are used such as one or two pitch periods and large errors due to aliasing have not been noticed.

4.4.1 Application of Cepstral Analysis to Speech

Voiced speech, $s_v(n)$, extracted from a speech segment is the convolution of a waveform, $p(n)$, which is repetitive at the pitch frequency, and vocal tract response, $h_v(n)$,(§2.3.2) (multiplied by a window $w(n)$), and can be represented as(Rabiner and Gold, 1975);

$$s_v(n) = [p(n) \odot h_v(n)].w(n), \tag{4.19}$$

Since $w(n)$ is generally a smooth sequence,4.19 can be simplified to the approximate form,

$$\begin{aligned} &\simeq [p(n).w(n)] \odot h_v(n), \\ &= p_w(n) \odot h_v(n). \end{aligned} \tag{4.20}$$

To separate the vocal tract response from the glottal waveform (assuming voiced speech produced from an impulsive repetitive glottal excitation) and hence determine the speech information the cepstrum can be invoked because $p_w(n)$ contributes a harmonic structure to the log spectrum, whereas $h_v(n)$ tends to contribute a smooth envelope. These two components therefore fall in different regions of quefrency in the cepstral domain. The pitch component of the cepstrum tends to consist of a narrow peak at a delay equal to the pitch period. The vocal tract component, by contrast, is smoothly varying in frequency so in the cepstral domain consists only of low quefrency components. Thus the cepstrum of (4.20) is written as the inverse Fourier transform of the log of the above components in the frequency domain, and for the continuous case,

$$c(t) = F^{-1}[log|P_w(f)| + log|H_v(f)|], \tag{4.21}$$

where F^{-1} represents the inverse Fourier transform. The cepstrum of the vocal tract component can be approximated by retaining only that portion of $c(t)$ for which $t < T_0$ where T_0 is the pitch period. Using a low-time window which is not too short will preserve the contribution due to $log|H_v(f)|$, spanning only $log|H_v(f)|$ and rejecting $log|P_w(f)|$.

A second method of calculating the cepstral coefficients is from the set of filter coefficients representing the vocal tract or LPC coefficients, provided they represent a stable filter. This is possible because the LPC filter $A(z)$ (where $A(z)$ is derived from the LPC coefficients such that $A(z) = \frac{1}{1-\sum_{k=1}^{P} a_k z^{-k}}$) has all of its roots inside the unit circle of the z-domain, implying that $log[A(1/z)]$ is analytic inside the unit circle and can be represented by a Taylor series (Atal and Hanauer, 1971). The Taylor series can be written in terms of the log magnitude spectrum,

$$log|H(z)| = C(z) = \sum_{n=-\infty}^{\infty} c_n z^{-n}, \tag{4.22}$$

where $z^{-n} = e^{-jwn}$ and c_n are the amplitudes at the nth sampling instant $t = nT$, of the inverse z-transform of $C(z)$. Replacing $H(z)$ in 4.22 by the linear prediction

formula (Atal and Hanauer, 1971),

$$H(z) = \frac{1}{1 - \sum_{k=1}^{p} a_k z^{-k}}, \tag{4.23}$$

and taking derivatives on both sides with respect to z^{-1} gives,

$$\frac{d}{dz^{-1}} log[\frac{1}{\{1 - \sum_{k=1}^{p} a_k z^{-k}\}}] = \frac{d}{dz^{-1}} \sum_{n=1}^{\infty} c_n z^{-n}, \tag{4.24}$$

which can be simplified to

$$\frac{[\sum_{k=1}^{p} k a_k z^{-k+1}]}{[1 - \sum_{k=1}^{p} a_k z^{-k}]} = \sum_{n=1}^{\infty} n c_n z^{-n+1}. \tag{4.25}$$

Terms can be equated to give the relationship between LPC coefficients and cepstral coefficients,

$$\begin{aligned} c_1 &= a_1 \\ c_n &= \textstyle\sum_{k=1}^{n-1}(1 - k/n) a_k c_{n-k} + a_n, \quad 1 < n < p \\ \text{and} & \\ c_n &= \textstyle\sum_{k=1}^{n-1}(1 - k/n) a_k c_{n-k}, \qquad n \geq p. \end{aligned} \tag{4.26}$$

Also note that c_0 can be calculated and is equal to $ln(G)$ where G is a gain factor representing the power of the system described by equation 4.5. Rather than calculating the cepstral coefficients from the Fourier transform, it is more efficient to calculate them from the LPCs using equation (4.26) This is the method employed in this thesis and is the usual method with recognition systems.

The number of coefficients chosen determines the length of the Taylor series representing the log spectrum. By omitting the higher order terms cepstral processing can therefore produce an equivalent smoothed version of the LPC spectrum, with smoothness dependent on the number of coefficients chosen (typically ten are calculated).

It is worthwhile here to note that due to the symmetry of the power spectrum, (i.e. $c_n = c_{-n}$), (4.22) can be rewritten as a pure cosine series expansion,

$$log|H(w,t)| = 2 \sum_{n=1}^{\infty} c_n(t) cos(nw) + c_0(t). \tag{4.27}$$

The cosine series expansion can be further approximated by a finite-term summation, called the discrete cosine transform(DCT). Zelinsky and Noll (1977) showed that the DCT basis functions have a close resemblance to the eigenvectors of the optimal Karhunen-Loève transform (also known as principle component factoring)(Cooley and Lohnes, 1971). The DCT representation of a short-time spectrum therefore approximates the optimal orthogonal Karhunen-Loève decomposition showing that the cepstral coefficients produce an optimal set of coefficients representing the speech samples. This optimality can also be observed by examining the covariance matrix of the cepstral coefficients which is diagonal dominant (Soong and Rosenberg, 1988) showing that there is little interaction between the dimensions.

Cepstrums first appeared in the early 1800s in the classical work of Poisson(1823), Schwarz(1872), Szegö(1915) and Kolmogorov(1939). This initial work was in the field of geophysical signal processing examining seismographs and echoes. In the speech field cepstral coefficients are widely used for word and speech recognition.

Cepstral coefficients were tested as recognition parameters in the early 1970s (Ichikawa *et al.*, 1973) giving 100% accuracy, an accuracy of over 20% better than LPC coefficients (§4.3). Many studies have given recognition results for cepstral coefficients in the 90% region and usually around 98 to 99%, (Juang *et al.*, 1987; Rabiner *et al.*, 1989; Furui, 1986). Brown and Rabiner (1982) compared cepstral coefficients and LPC coefficients using Euclidean (§6.2) and log likelihood distance (§6.4) measures respectively. The cepstral distance was consistently better for the speakers and vocabulary used. Tohkura (1987) also tested cepstral coefficients against LPC coefficients. Recognition with cepstral coefficients used both Euclidean and weighted Euclidean distance (Mahalanobis distance) (§6.2) while for LPC coefficients a log likelihood distance was used. The weighted distance measure with cepstral coefficients gave improved performance over both other distances. Juang (1987) claimed increased performance with liftering (§6.2) cepstral coefficients. Recognition errors as low as 1% for speaker independent recognition of the digits was claimed. The SPHINX speech recognition (Lee *et al.*, 1990) also used cepstral coefficients in a HMM large vocabulary speaker independent continuous speech recogniser. Accuracy as high as 96% for a 997-word task was reported.

4.5 DYNAMIC CEPSTRAL COEFFICIENTS (DCEP)

Spectral transitions as well as instantaneous spectral features are believed to be important for sound recognition (Soong and Rosenberg, 1988; Furui, 1988; Hunt and Lefebvre, 1989). Dynamic spectral feature analysis is still in its infancy and only a few researchers are using this information for word recognition, although it has been used more widely as a speaker recognition tool (Furui, 1981; Soong and Rosenberg, 1988). Simple first order finite differences are far too noisy to be used as dynamic information so the method used in this study follows that of Soong and Rosenberg (1988). Using the LPC-based cepstral coefficients ($c_n(t)$), orthogonal polynomials which characterise the time trajectories of the cepstral coefficients over a finite number (2K+1) of fixed length frames are calculated. For a first order polynomial two coefficients are calculated. The zeroth order (ZCEP) or constant term is given by

$$\hat{c}_n(t) = \frac{\sum_{k=-K}^{K} h_k c_n(t+k)}{\sum_{k=-K}^{K} h_k}, \tag{4.28}$$

for a number of calculated coefficients $n = 1..N$, where h_k is a symmetric window function of length (2K+1) frames. The first order (FCEP) orthogonal polynomial coefficient, or spectral slope is

$$\Delta c_n(t) = \frac{\sum_{k=-K}^{K} k h_k c_n(t+k)}{\sum_{k=-K}^{K} h_k k^2}. \tag{4.29}$$

The window length, referenced as $2K+1$, can have a significant effect on the ability of this measure to model the movements in time. A window of reasonable length has to be used to ensure a smooth transition of the coefficients from one frame to the next. Soong and Rosenberg(1988) performed recognition tests with different window lengths, finding that seven frames per window was optimal. In contrast Furui(1986) claims that the dynamic coefficients extracted from nine frame intervals gives only slightly better recognition performance than the instantaneous cepstrum coefficients. Svendsen *et al* (1989) improved their recognition scheme by adding dynamic cepstral information in conjunction with instantaneous cepstral data. Recognition accuracy improved from 84.7% to 98%. Nishimura(1989) also tested instantaneous and dynamic features claiming higher robustness for transitional over instantaneous features for HMM recognition.

When testing female speakers, who gave high recognition errors with instantaneous features, a drop in recognition error of up to 75% was achieved when dynamic features were used.

The highest accuracies have been obtained by employing a combination of both features. Furui (1989) proposed the use of instantaneous and transitional cepstral features within a vector quantisation framework for isolated word recognition. In all cases the combination of instantaneous and transitional data proved more accurate than using instantaneous data alone.

4.6 PERCEPTUAL LINEAR PREDICTORS (PLP)

The human auditory system is an excellent speech recogniser. If computer models can adequately reflect this system by reproducing the transformations that take place in the ear then the resulting processing techniques should give performance superior to more simplistic approaches.

The problems of modelling the human ear are immense due to its phenomenal properties. Consisting of three sections (outer, middle and inner) the ear forms a complicated frequency analyser of high selectivity. The ear is capable of responding to frequencies from 20Hz to 20kHz and to detect a sound which displaces the eardrum only one-tenth of an Angstrom.

Although not all the transformations achieved by the ear are known, the psychophysical properties of the ear have been studied by a number of researchers (Shaw, 1980; Allen, 1985). The effects are those which can be measured in terms of a listener's subjective characteristics. Experimental data from many listeners can agree with surprising accuracy on many subjective parameters, such as just noticeable differences for frequency and intensity, and loudness (Kinsler *et al.*, 1982). From these studies it is clear that the ear performs a frequency analysis but with nonlinearities, such as saturation at high stimulus levels, and dynamic effects, such as adaptation (Seneff, 1986). To model the ear effectively relationships must be found that relate the subjective quantities to physical parameters.

Models of human perception can be divided into two categories; those that model the physiology of the human hearing and those that model psychoacoustic phenomena in terms of human responses to acoustic phenomena. The first to propose a model based on physiology for speech recognition was Cannon in 1968. Since then many physiological models have been proposed, (Lyon and Dyer, 1986; Seneff, 1986; Hunt and Lefebvre, 1987; Ghitza, 1988; Gramss and Strube, 1990). These systems are highly complicated and are unsuitable for speech recognition. Psychoacoustic models have been reported by Itahashi and Yokoyama(1976), Hermansky(1985) and Junqua(1991). Hermansky's model, first described in 1985, has several advantages over other systems for real-time processing. Firstly the modelling of the auditory spectrum with a low order model, using only a fifth order model, makes the calculation of Hermansky's PLPs faster than LPC calculation which usually requires that 10 or more coefficients be calculated (Hermansky *et al.*, 1985). Secondly high recognition results have been claimed for speaker dependent and independent tests (Hermansky *et al.*, 1985). Thirdly this method has also produced high recognition accuracies for cross-speaker word recognition, such as between male and female speakers (Hermansky *et al.*, 1986). For these reasons it was decided to test this model in our system.

Hermansky's auditory spectrum, see Fig. 4.9, is produced by critical band filtering, followed by equal loudness curve pre-emphasis and intensity-loudness conversion. A fifth order all-pole model (LPC) is then used to extract at least two major peaks from the auditory spectrum shape.

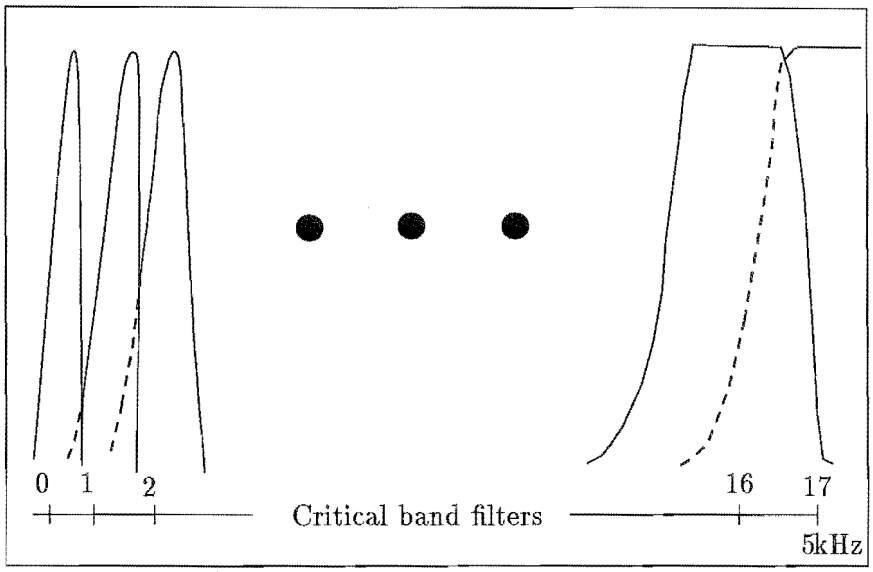


Figure 4.7. Critical band filters used to model the ear's response to frequency input.

The critical band filtering follows from the representation of the ear as a collection of parallel filters (Fallside and Woods, 1983), each with a different bandwidth as shown in Fig. 4.7. The bandwidth of each filter is nearly 1/3 octave for filters with centre frequencies above 400Hz. For Hermansky's representation 18 critical band filters are specified with their centre frequencies equally spaced in the Bark domain. The power within each critical band filter is calculated by summing the weighted short-term power spectra $P(w)$, derived by the FFT algorithm. For the frequency range $0 \leq f \leq 5$kHz $(0 \leq w \leq \pi)$ the critical bands cover the range $0 \leq \omega \leq 16.9$Bark. The Bark values, ω are calculated from the frequencies, f using the natural log function ln such that,

$$\omega = 6ln\left[\frac{f}{600} + \sqrt{(\frac{f}{600})^2 + 1}\right] \quad 0 \leq f \leq 5000\text{Hz} \tag{4.30}$$

The critical band centre frequencies, ω_k are spaced such that for the kth critical band the center frequency is,

$$\omega_k = 0.99k. \tag{4.31}$$

Each critical band is weighted by a function that determines its shape. The weighting is such that the band shape is asymmetric being less steep towards the lower frequencies. The critical band weighting function which was used by Hermansky is dependent on the bark frequency within the band and specifies the band shape. For the kth band its weighting is given by,

$$C_k = \begin{cases} 10^{\omega-\omega_k+0.5} & \omega \leq \omega_k - 0.5 \\ 1 & \omega_k - 0.5 < \omega < \omega_k + 0.5 \\ 10^{-2.5(\omega-\omega_k-0.5)} & \omega \geq \omega_k + 0.5. \end{cases} \tag{4.32}$$

Equal loudness levels can be found by experiments which gauge when two tones of dissimilar frequencies appear to be equally loud. The loudness level is very dependent on the frequencies of the tones, as shown in Fig. 4.8. High and low frequency tones require greater intensity to sound as loud as those in the midfrequency range. Equal

loudness is simulated by Hermansky *et al*(1985) by pre-emphasis of the speech and then by weighting with,

$$E(w) = 1.151\sqrt{\frac{(w^2 + 1440000)w^2}{(w^2 + 160000)(w^2 + 9610000)}}. \tag{4.33}$$

For a listener, if two or more tones are sounded simultaneously, each with an intensity I, the total loudness (N) depends on whether they lie within a single critical band. Tones which lie within a critical band have a loudness given by the sum of their intensities,

$$Loudness(Sone) = 460F(f)(\sum_i I_i)^{\frac{1}{3}}, \tag{4.34}$$

where $F(f)$ is a parameter, empirically determined, and dependent on frequency. Hermansky uses this 1/3 power law to calculate the outputs, Q_k of each of the bands such that,

$$Q_k = [E(w)\int_0^{\pi} C_k(w).P(w)dw]^{\frac{1}{3}} \quad 1 \leq k \leq 17, \tag{4.35}$$

where $P(w)$ is the short-time power spectrum of the speech derived by the FFT algorithm.

Hermansky(1985) notes that the approximation of the critical band weightings given in (4.32) is not very accurate close to the zero or Nyquist (5kHz) frequencies. Because of these inaccuracies the zero band value is not calculated. To obtain a value for the zeroth band Hermansky equates the zeroth frequency band value to the first frequency band, resulting in 18 critical band outputs.

To obtain a LPC representation, the 18 filter band outputs are interpolated to a 129 point representation. The interpolation between filter band outputs is calculated as,

$$Q(w) = [Q_k + [(Q_{k+1} - Q_k)/(\omega_{k+1} - \omega_k)](\omega - \omega_k)] \tag{4.36}$$

$$\omega_k \leq \omega \leq \omega_{k+1}, \quad k = 0, 1..16 \tag{4.37}$$

which is used to obtain a 129 point representation of the critical band spectral envelope in the loudness domain. To approximate the interpolated auditory spectrum by a 5th order all-pole model the inverse discrete Fourier transform(DFT) is used to find six terms of the autocorrelation function. A fifth order LPC model is then calculated using the Durbin-Levinson algorithm outputting coefficients Hermansky calls perceptual linear predictors (PLPs) The calculation of PLP coefficients is drawn in block diagram form in Fig. 4.9.

Hermansky used PLPs as features in a DTW matching algorithm, (§5.1). Recognition accuracies of 90-95% were quoted for speaker independent recognition of the digit vocabulary using only two templates per word. Cross speaker recognition, where male-female recognition and male-male recognition is performed, gave better recognition accuracy over LPC (Hermansky *et al.*, 1986).

Hanson *et al* (1988) compared PLP and cepstral coefficients for word recognition. For speaker dependent recognition they found that the fifth order PLP representation gave lower accuracy than the standard cepstral representation. Their speaker-dependent recognition results indicated that some phonetically relevant information is discarded by the PLP analysis because of the reduction in detailed spectral distinction caused by using a fifth order PLP. When Hanson *et al*(1988) increased PLP order to 14, recognition accuracies were also increased (and were above that for cepstral coefficients) for all male speakers. However the female speakers' PLP recognition accuracy was still lower than that of cepstral coefficients.

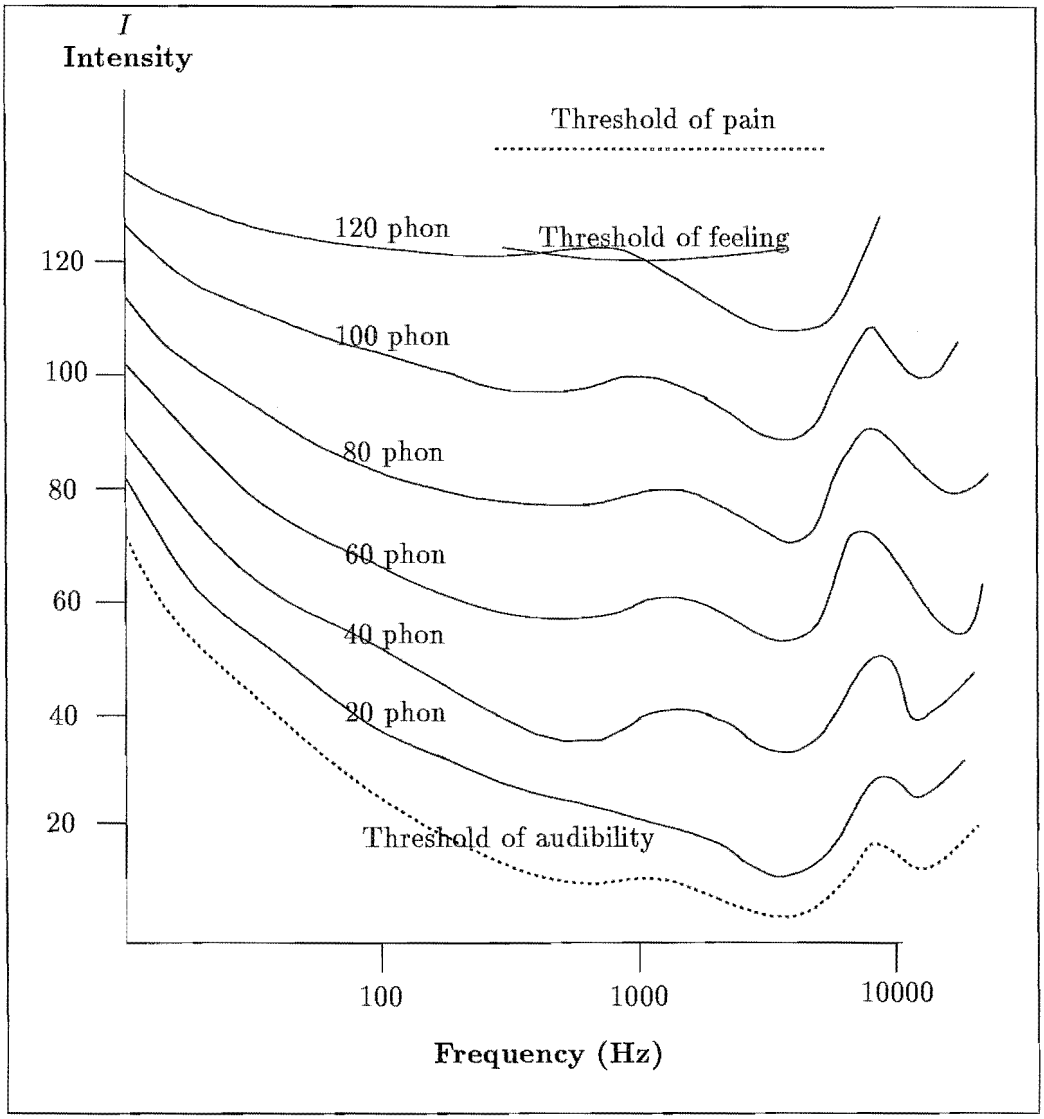


Figure 4.8. Equal loudness curve showing equivalence between intensity of sound and the loudness heard with respect to the frequency of a tone.

4.7 SUMMARY

From the examination of different speech feature representations it appears that a method which involves the modelling of the frequency content of the speech may be more worthwhile than a time base representation. Simple temporal representations, such as ZX and RMS, do not adequately model the important information required for speech recognition, although RMS and ZX may be adequate to distinguish word variations within a small vocabulary. Further, methods which take into account how coefficients change with time, such as transitional coefficients, appear to also give high recognition accuracies. From these studies it seems important to represent the transitional as well as the instantaneous speech information. Perceptual features, which model the way the human listener perceives speech have also given highly accurate results. Although many of the feature representations and their performances as recognition parameters have been compared in the literature no comparisons have been made of all these features under standard conditions. A comprehensive comparison of

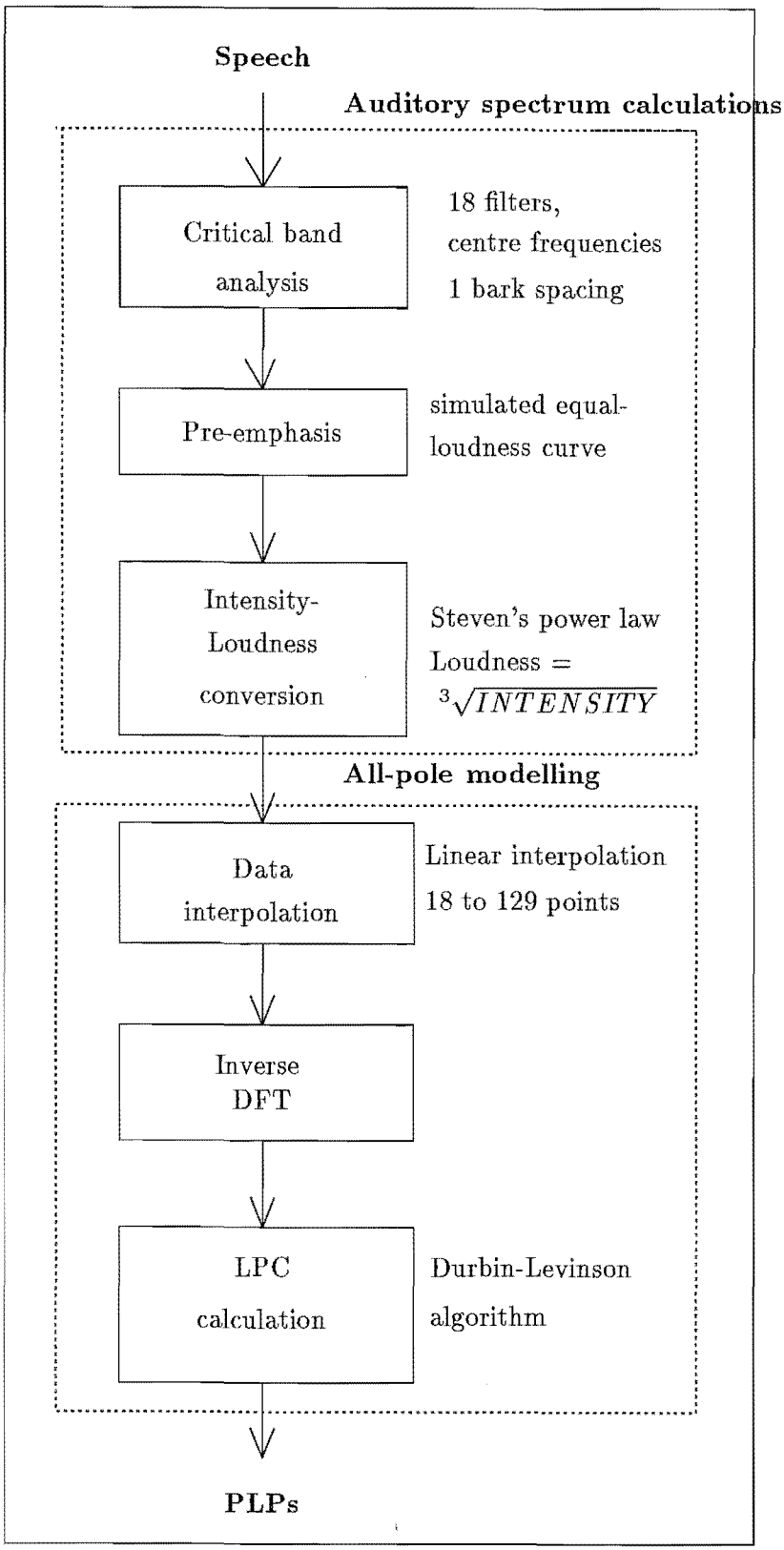


Figure 4.9. Block diagram of calculation steps for perceptual linear predictors (PLPs).

all these features is undertaken and discussed fully in Chapter 8.

Chapter 5

METHODS OF RECOGNITION BY DYNAMIC ALIGNMENT

A major problem associated with recognition of speech is the variability of the relative duration of component sounds when words are spoken at different rates. When comparing words of different lengths some form of time-normalisation is required to account for both the overall word variability and the within word variability. Initial researchers used linear normalisation techniques, but these simple techniques were found to be inadequate for dealing with the inherently non-linear distortions (Sakoe and Chiba, 1978). During the early 1970s many methods were proposed for non-linear matching. Two methods can be broadly identified as most successful - dynamic time warping(DTW) and hidden Markov modelling (HMM). These two methods are discussed in sections 5.1 and 5.2 respectively.

Another recently emerging method proving to be useful for word recognition is that of neural nets. Although this method is showing promise it is still being developed and its abilities and limitations are not well known. In the work presented in this thesis, where a comparison is being undertaken of many features types, it would be impractical to use a method in which the fundamental limitations are not clearly understood. Because of this neural networks are not considered in this thesis, further discussions can be found in Lippmann, 1987.

5.1 DYNAMIC TIME WARPING (DTW)

DTW uses the dynamic programming (DP) optimisation technique first proposed by Bellman (1957) (refer §5.1.3). The technique of DP can be applied to the recognition of speech to produce the DTW process because the speech systems have the following features, which are common to systems that can be optimally solved using DP (Bellman, 1957);

- The physical system can be characterized at any stage by a small set of parameters, the state variables.
- At each stage of the process there is a choice of a number of decisions.
- The effect of a decision is a transformation of the state variables.
- The past history of the system is of no importance in determining future actions.
- The purpose of the process is to maximize some function of the state variables.

DP applied to time varying signals produces a time-normalisation procedure known as DTW. The time-normalisation effect caused by DP eliminates the timing differences between two speech patterns by warping the time-axis of one so that the maximum coincidence is obtained with the other. The maximum coincidence is calculated by minimising a distance between the two words. By minimally optimising the distance

between the words timing fluctuations are produced which are modelled by a non-linear warping function. Because one template is matched to another during the DTW process, the process belongs to the class of recognisers which are based on template matching.

5.1.1 Dynamic Time Warping for Isolated Word Recognition

Two speech sequences (or words) of different length, expanded as a series of extracted features, can be represented as,

$$R_M = \{r_1, r_2, r_3 ... r_M\} \tag{5.1}$$

$$T_N = \{t_1, t_2, t_3 ... t_N\}, \tag{5.2}$$

where R_M, a stored template, is the *reference* word of length M containing extracted feature vectors $r_1, r_2, ...r_M$ and T_N is an input *test* word of length N containing extracted feature vectors $t_1, t_2 ... t_N$, whose similarity to R_M is to be quantified. Where calculations are required over a shorter set of features, say m_k where $m_k < M$ or to n_k where $n_k < N$ then the speech sequences to that length can be represented as

$$R_m k = \{r_1, r_2, r_3 ... r_{mk}\} \tag{5.3}$$

$$T_n k = \{t_1, t_2, t_3 ... t_{nk}\}. \tag{5.4}$$

Both R_M and T_N (or $R_m k$ and $T_n k$) are described by sequences of multidimensional feature vectors, with each feature vector represented by r_i or t_i, such as linear prediction coefficients. Fig. 5.1 shows the stored speech pattern, R_M, and the test pattern T_N expanded along the m and n axes respectively. In this case, DTW is matching an unknown pattern, the test pattern T_N, to a reference pattern, R_M. The DTW problem is to find the optimal warping path which minimises the sum of *distances* measured between the two patterns. Each distance is calculated at a frame of the speech which is represented by a feature vector and is equivalent to an intersection point on Fig. 5.1. By finding the minimum distance path through these grid points an optimal path can be found. This optimal path can be represented geometrically as the mapping of axis n on to axis m by a warping function w,

$$m = w(n). \tag{5.5}$$

Since, in general, the warping function does not coincide with a grid point for all values of n and m it is convenient to introduce an index, k for those grid points which lie on the warping function. These grid points are denoted (n_k, m_k) where

$$m_k = w(n_k), k = 1 ... K \tag{5.6}$$

and k is the index along the warping path and K is the number of distances which are summed to find the optimal global distance between the test and reference words. For the most general type of warping procedure (shown in Fig. 5.1) where the warping path starts at position (1,1) and finishes at (N, M), $k = 1$ and $k = K$ correspond to the points $(1, 1)$ and (N, M) respectively.

Although the DTW optimal warping path can be found by searching over the entire n, m axes this is computationally inefficient. For speech, constraints can be placed on the movement allowed for the warping path thus limiting the number of calculations required to find the optimal path.

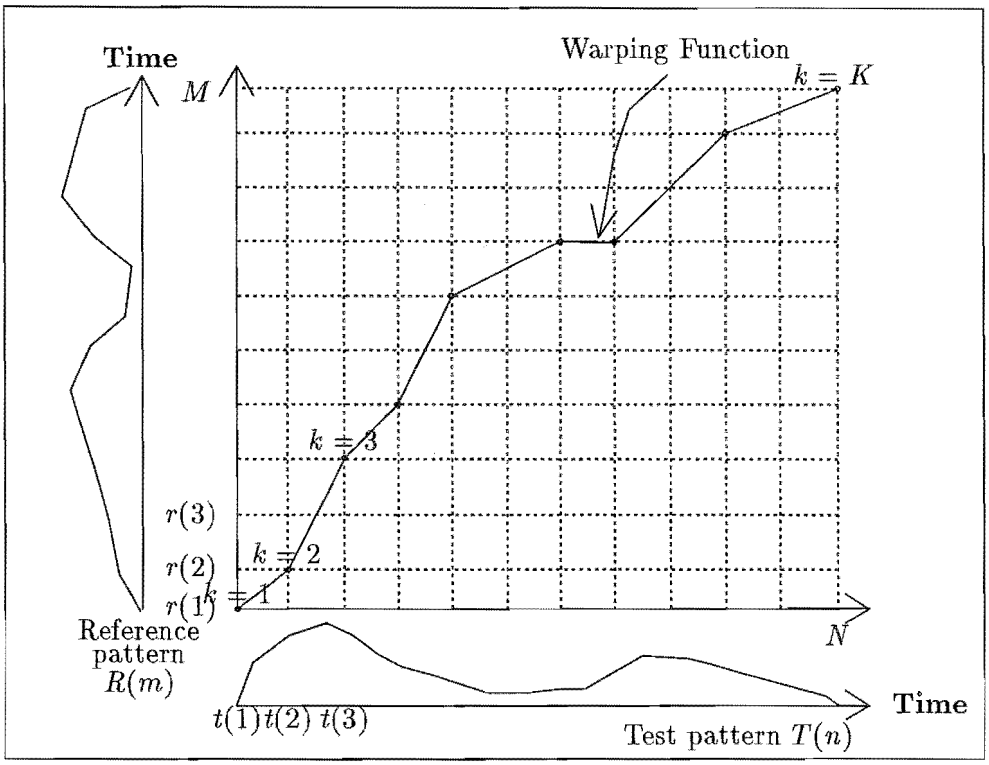


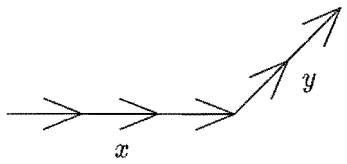
Figure 5.1. Illustration of a non-linear mapping between two speech patterns, which could have been produced using DTW. The mapping is warping a stored speech pattern R_M, known as the reference template, to a test pattern T_N. R_M and T_N are shown as continuous representations of discrete functions of feature vectors. At its most general, the DTW algorithm finds the path corresponding to a minimum accumulated distance between the frames of data, where each frame of data corresponds to the intersections of the dotted lines (for position n and m this corresponds to the data frames $t(n)$ and $r(m)$). At each data frame a local distance is calculated. The warping function, $w(n_k)$, corresponds to the least cost of the accumulated distances, where the accumulated distance is found by summing the local distances along every possible path through the grid from beginning to ending. The warping function $w(n_k)$ is indexed by k where k counts the number of minimum distances summed to find the global distance between the reference and test patterns.

5.1.2 Warping Function Restrictions

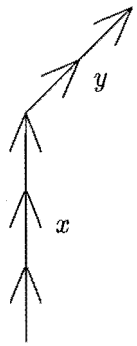
The general DTW method, discussed in §5.1.1, can be optimised by introducing constraints based on known speech pattern structures. Such speech structures that are being modelled by the DTW method include continuity, monotonicity (relative timing restrictions), and acoustic parameter transition speed (Sakoe and Chiba, 1978). Speech structures may also affect what occurs at the DTW boundaries. These structures impose constraints and conditions on the DTW algorithm which will be discussed in the following sections.

5.1.2.1 Slope Constraints.

To prevent meaningless comparisons between a very long speech pattern and a very short speech pattern the warping function cannot be allowed to be too steep or too shallow. The slope constraint is realised as a restriction on the possible movement of consecutive points on the warping function. Thus if the warping function moves a



(a) **Minimum Slope**



(b) **Maximum Slope**

Figure 5.2. Slope constraint placed on warping function reducing meaningless comparisons. The constraint illustrated reduces movement of the warping function $w(n_k)$ so that if it moves a distance x in the m (or n) direction then it cannot move any further in that direction until it travels a distance y along the $m = n$ direction. Illustration shows (a)minimum and (b)maximum movement of warping function.

distance x in the direction of m (or n), then the warping function cannot travel any more in that direction until it moves a distance y along the line $m = n$ as illustrated in Fig. 5.2.

The parameter used to constrain the slope is measured as,

$$\begin{aligned} P &= \frac{\text{movement allowed in the m=n direction}}{\text{movement allowed in the m (or n) direction}} \\ &= y/x, \end{aligned} \tag{5.7}$$

The larger P is, the more rigidly the warping function slope is restricted. When $P = 0$ there is no restriction on the warping function. When $P = \infty$, that is x=0 the warping function is restricted to the diagonal line $m = n$ forcing only a linear warp.

5.1.2.2 Continuity Constraints

Local continuity constraints restrict excessive compression or expansion of the time scales, restricting the warping path movement from one grid point to another. For example one type of constraint could restrict movement such that,

$$\begin{aligned} &m_k - m_{k-1} \leq 1 \\ &\text{and} \\ &n_k - n_{k-1} \leq 1, \end{aligned} \tag{5.8}$$

For such a case Fig. 5.3 shows allowable path movements in a *symmetric* form where the movement in the n and m directions is the same. However more complex

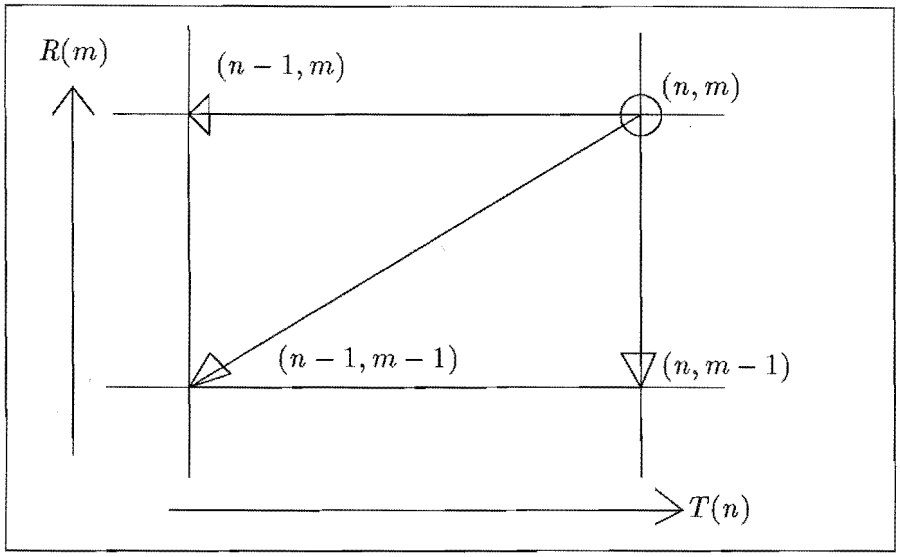


Figure 5.3. Illustration of a general set of local path constraints affecting warping path movements to reach point (n, m). The warping path is shown as if calculations are done during back-propagation, however in reality the calculations are performed in the forward direction towards (n, m). Warping path movements of this kind are known as *symmetric* because they move equally in the m and n directions.

constraints can be allowed, such as those imposed by Itakura(1974), (as illustrated in Fig. 5.4) giving *asymmetric* path movements where the movement in the n and m directions is not the same and constraining warping function movements such that,

$$w(n_k) - w(n_{k-1}) = \begin{cases} 0, 1, 2 & if \quad w(n_{k-1}) \neq w(n_{k-2}) \\ 1, 2 & if \quad w(n_{k-1}) = w(n_{k-2}). \end{cases} \tag{5.9}$$

The permissible warping movements for arriving at n and m as specified by equation 5.9 are shown in Fig 5.4.

Global continuity constraints are affected by the local path constraints. Because of the local path constraints, certain parts of the (n, m) plane are excluded from the region in which the optimal warping path can lie. Using a local path constraint such as that of equation (5.9) limits the overall slope of the time alignment contour to be between $1/2$ and 2. With slope constraints of 2 and $1/2$, such that the warping function cannot move with a slope greater than 2 or less than $1/2$, the path will have a global constraint shown in Fig. 5.5, a parallelogram bounded by lines of slope 2 and slope $1/2$. The slope constraint of $1/2$ is determined by the condition that the optimal path cannot be flat ($w(n_k) - w(n_{k-1}) = 0$) for two consecutive frames, and the slope constraint of 2 is determined by the condition that no path to the grid point (n, m) can come from any grid point lower than $(n - 1, m - 2)$. These methods are often known as 2-to-1 methods and are here referenced as 21 methods.

5.1.2.3 Monotonic Condition

Monotonic or relative timing conditions restricts the order in the warping by taking into consideration the the timing in the speech such that

$$n_{k-1} \leq n_k, \tag{5.10}$$

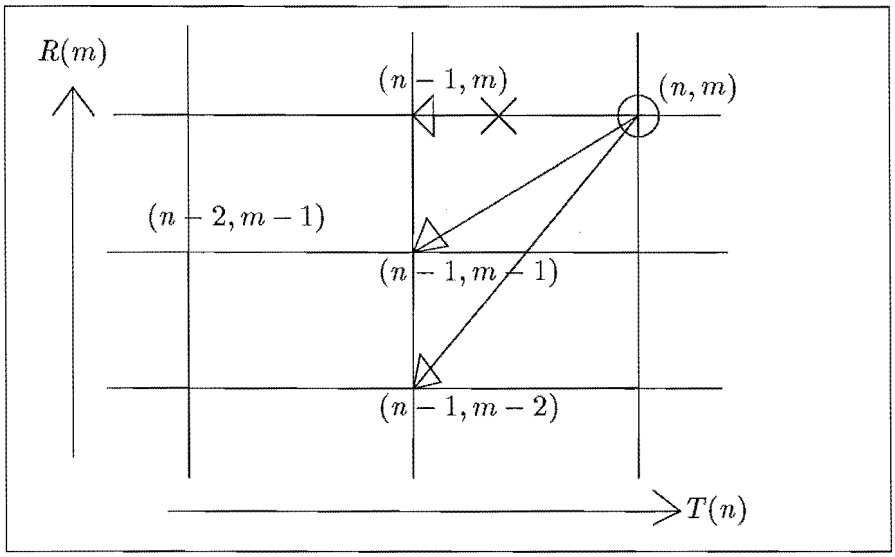


Figure 5.4. Illustration of the Itakura local path constraints affecting warping path movements to reach point (n, m). The warping path is shown as if calculations are done during back-propagation, however in reality the calculations are performed in the forward direction, towards (n, m). Note that the movement from point $(n-1, m)$ can only be considered if this type path movement was not taken in the previous warping path calculation, and this is signified by a crossed segment. Warping path movements of this kind are known as *asymmetric* because movement is greater in the m direction than the n direction.

and

$$m_{k-1} \leq m_k. \tag{5.11}$$

5.1.2.4 Local Weighting Constraints

Weightings along the search paths of the warping function can bias the transitions of the warping function towards particular path movements. The weighting on the warping path accounts for the probability distribution of the acoustic transitions of the speech and is equivalent to knowing *apriori* how much the sound is expected to change from one frame to the next. Where weightings are used the frames of data should be weighted so that the warping function is coaxed along the diagonal. Fig. 5.6 shows a series of weighting functions used in a general DTW scheme. Obviously these weightings can lead to a preference for some paths over others. Weighting of type (II) will lead to a preference for the longer path while type (III) weighting will give preference to the shorter path.

5.1.2.5 Boundary conditions.

Two methods of applying boundary conditions exist; the constrained endpoint(CE) and the unconstrained endpoint(UE).

The CE method sets the beginning and ending of the warping function at exact positions; the beginning at the first frame of both the test and reference words and the ending at the last frame of both the test and reference words, that is the start point of the warping function is set at $(1, 1)$ and the end point is set at (N, M). This method is primarily used when precisely determined endpoints for both the reference and test patterns can be found.

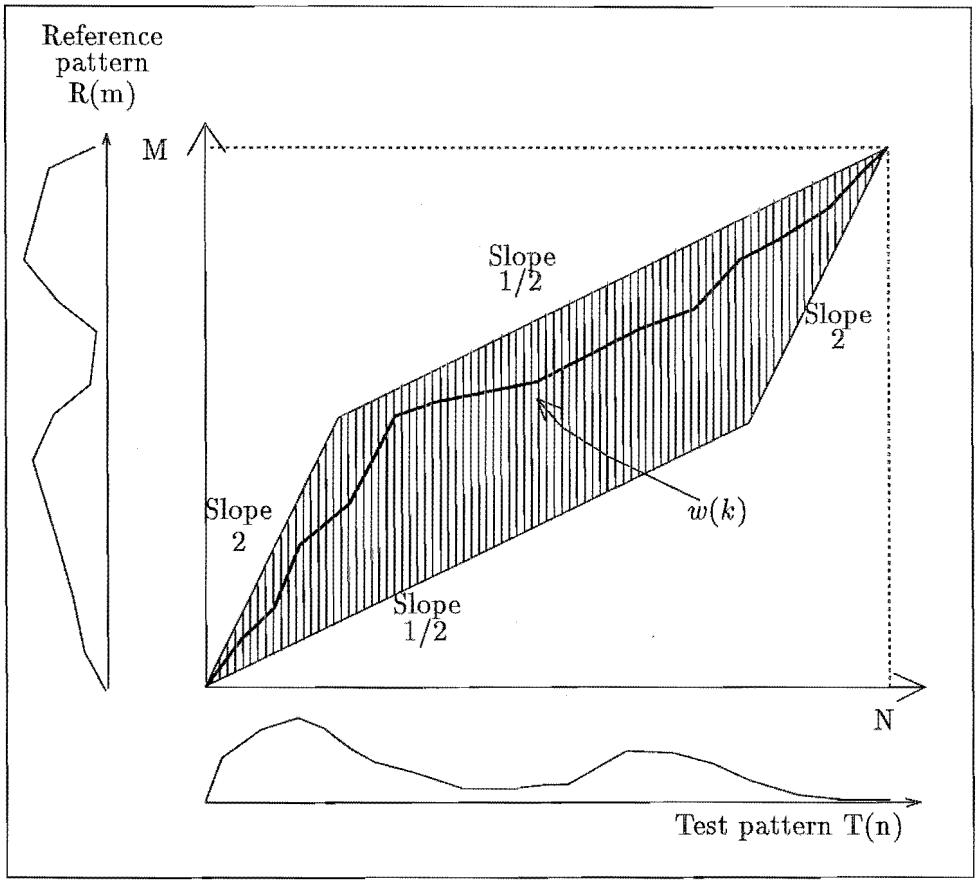


Figure 5.5. Illustration of global constraints on the range allowed by the warping function. The global range restricts the length of test and reference patterns that can be warped together, in this case test length cannot be greater than twice the length of the reference or shorter than one-half the length of the reference. The hatched area shows the region of possible path movements allowed for the warping path.

The UE method relaxes the endpoint constraint such that,

$$1 \leq w(1) \leq 1 + \sigma, \tag{5.12}$$

is the beginning condition and

$$M - \sigma \leq w(N) \leq M, \tag{5.13}$$

is the ending condition. σ is an offset parameter dependent on the amount of endpoint relaxation allowed. When $\sigma = 0$ the UE method becomes the same as the CE method.

5.1.3 Distance Measure.

Arguably the most important factor relating to the DTW algorithm (in the context of word recognition) is the distance function whose minimisation defines the optimal warping path and is further discussed in Chapter 6. A general form for such a distance function is,

$$D(T_K, R_K) = \frac{\sum_{k=1}^{K} d(t_{nk}, r_{mk}) W'(k)}{K}, \tag{5.14}$$

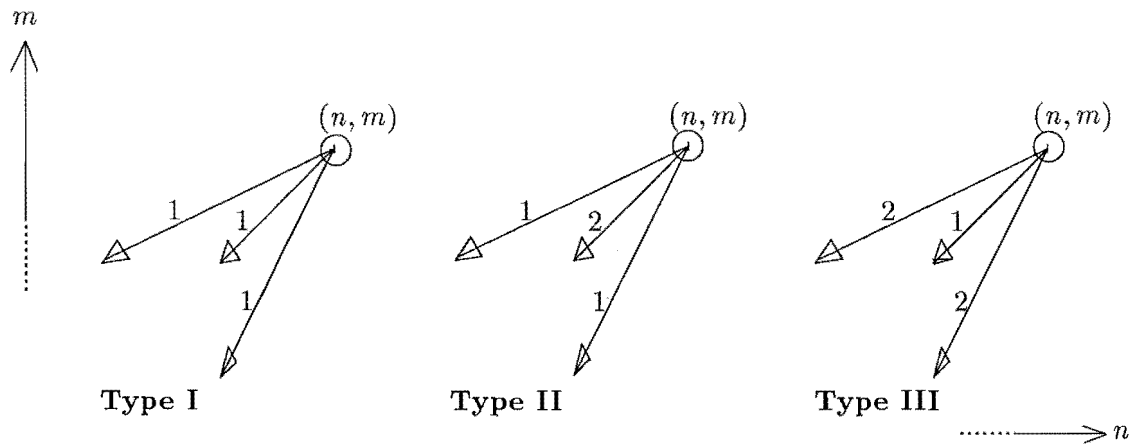


Figure 5.6. Illustration of the type of weighting applied to DTW warping path movement. Warping path movement is shown along the (n, m) axes. Weightings for each path movement are given on the slope of the vectors. Type I weightings gives equal weighting for all warping path transitions. Type II weightings weight shorter warping path transitions greater than longer ones, giving preference to longer paths in the distance measure. Type III weightings weight longer warping path transitions greater than shorter ones, giving preference to shorter path movements in the distance measure.

where $D(T_K, R_K)$ is the average distance along the warping path and is calculated by accumulating the local distances $d(t_{nk}, r_{mk})$ and therefore depends on the warping function. $d(t_{nk}, r_{mk})$ is the local distance between frames t_{nk} and r_{mk}. The local distances depend only on the feature set used to describe the test and reference patterns and not on the warping function, although the test and reference data frames which are used for the distance calculation do depend on the warping function. $W'(k)$ is the weighting function which was discussed in §5.1.2.4. K is a normalisation factor determined by the constraint that the distance $D(T_K, R_K)$ be equivalent to the average local distance along the path, and hence is not affected by lengths of either the test or reference pattern. Normalising the distance in this way makes its value independent of the length of the warping path.

Although the optimal warping path is calculated by backtracking from the end of the path to the beginning, the total accumulated distance $D(T_K, R_K)$ can be calculated as the path is formed, from the beginning. This is true due to the *principle of optimality* outlined by Bellman(1957) which states *An optimal policy has the property that whatever the initial state and initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision* Thus the optimal path to the grid point (n, m) only depends on values of n' and m' such that $n' \leq n$ and $m' \leq m$ and is not dependent on grid points past n and m. This points to the basic idea of dynamic programming which is the replacement of the optimization of a function by the solution of a number of optimization problems, which are easier to solve (Minoux, 1986).

Using Bellman's principle it is possible to create a partial accumulated distance as $D(T_{nk}, R_{mk})$ along the best path from $(1, 1)$ to (n_k, m_k). Illustrating this using type III constraints shown in Fig. 5.6 and discussed in §5.1.2.4, this accumulated distance $D(T_{nk}, R_w(nk))$ at grid point (n_k, m_k), where n_k is the grid point corresponding to t_{nk}

and m_k is the grid point corresponding to r_{mk}, can be written as,

$$D(T_{nk}, R_{mk}) = D(n,m) = min \begin{cases} D(n-1, m-1) + d(n,m) \\ D(n-1, m-2) + 2d(n,m)) \\ D(n-2, m-1) + 2d(n,m)). \end{cases} \quad (5.15)$$

where $D(n-1, m-1)$ represents the partial accumulated distance calculated by accumulating the distance to and including grid point $(n-1, m-1)$, and $d(n,m)$ is the distance calculated at grid point (n,m). When the accumulated distance is calculated to the word endpoints constraints the minimised distance between the two words is known.

5.1.4 DTW Variants

Fig. 5.7 illustrates three versions of DTW algorithms. The first, called the constrained endpoint, 2-to-1 slope (CE21), is the general method which assumes perfect endpoint alignment. A variation on this method, the unconstrained endpoint 2-to-1 slope method (UE21) relaxes endpoint constraints as discussed previously in §5.1.2.5. A third variation, the unconstrained endpoint band method (UEB) also relaxes endpoint constraints but the warping path is constrained to lie within a band. The width of the band is kept at a constant value and is centered at the value m^* at which the minimum accumulated distance $(D(n-1, m^*))$ was calculated. A path constraint can therefore be written of the form,

$$m^*(n-1) - \alpha/2 \leq m(n) \leq m^*(n-1) + \alpha/2, \quad (5.16)$$

where $m(n)$ is the range on m over n in which an optimum warping path is searched for, $m^*(n-1)$ is the m index giving the minimum distance at $n-1$, and α is the width of the band.

5.1.5 Uses of DTW

Highly accurate dynamic time warping (DTW) word matching techniques have been reported from as early as 1974. These first experiments were on Japanese digit words, giving recognition accuracies as high as 99.8% (Itakura, 1974). Since then DTW has been one of the most widely used techniques for isolated word recognition. Modifications to DTW word recognition techniques has allowed connected and continuous word recognition systems to be developed based on DTW (Sakoe, 1979; Rabiner and Schmidt, 1980; Myers and Levinson, 1982; Ney, 1984). Other schemes which use DTW procedures aim to align natural and synthetic speech to improve the quality of synthetic speech (Hohne *et al.*, 1983) and to recognize words embedded in continuously spoken sentences (Christiansen and Rushforth, 1977).

5.1.5.1 Connected and continuous Recognition using DTW

Connected word recognition is the recognition of sentences where each individual word is spoken with slight pauses inbetween. The technology to recognize sentences is based on isolated word recognition because, to recognize the sentence of words, each individual word is recognised. *Continuous word recognition* is the recognition of sentences of words that are spoken continuously (without pausing). This type of recognition requires sentence meaning to be determined because it becomes too difficult to recognize each individual word. Connected word and continuous speech recognition systems require an integration of speech knowledge and application of heuristics to achieve accuracy. While connected word recognition methods have been extensively researched within the last

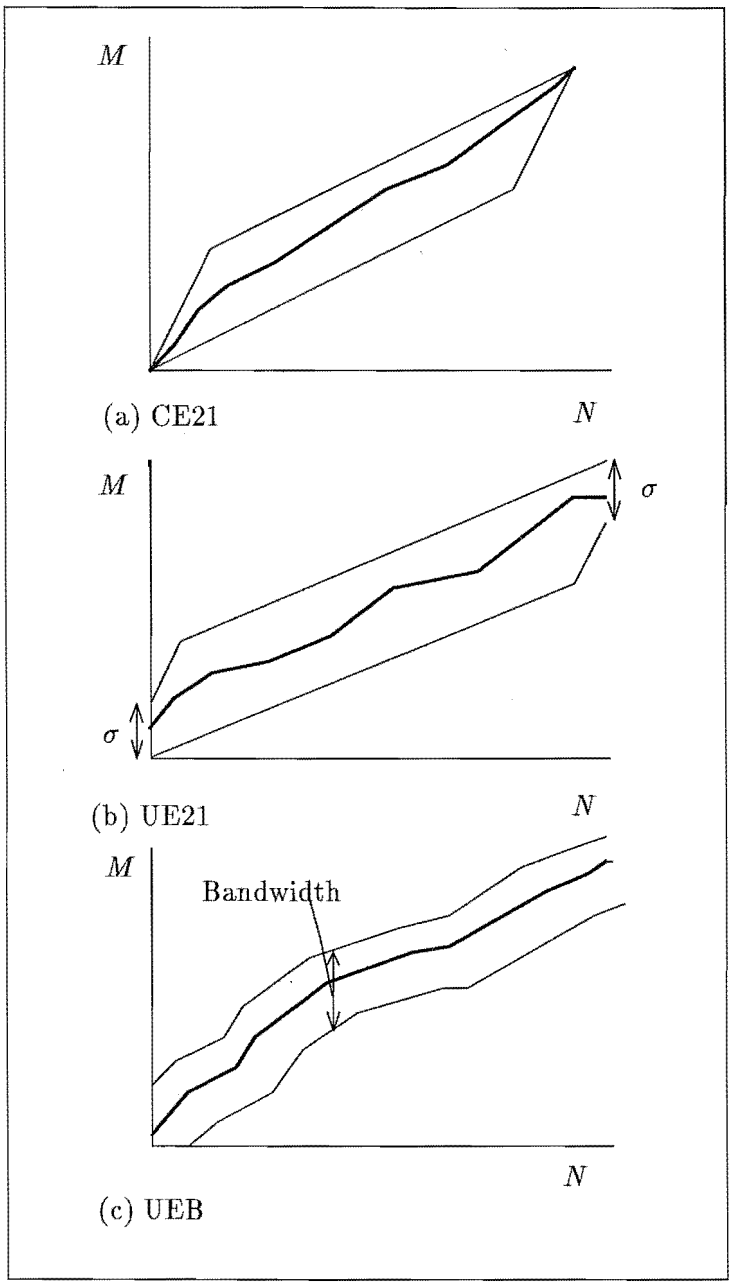


Figure 5.7. Variations of DTW schemes showing regions of possible path movements. (a) The constrained endpoint 2-to-1 method, (CE21), (b) the unconstrained endpoint 2-to-1 method, (UE21), (c) the unconstrained endpoint band method, (UEB). Note that the path movements in methods (a) and (b) are set prior to warping path calculation, while for method (c) path movements are dependent on the warping path during calculations.

few years, their practical use has been limited to small vocabularies and rigid syntactic structures. However these systems have been more successful than their continuous speech recognition counterparts.

The major algorithms reported in the literature for connected word recognition systems include the two-level (TL) (Sakoe, 1979), the level-building (LB) (Myers *et al.*, 1981) and the one-pass algorithm (OP) (Bridle *et al.*, 1982). Although all these algorithms are attempting to solve for the same global optimum, that is the minimum

accumulated distance between a test string and a series of reference templates, how this is achieved is unique for each system. This difference means that the level-building and two-level methods can not operate in real-time while the one-pass method can. Because a real-time method was required in this thesis only the one-pass method was seriously considered, however the following section gives a brief overview of all these three methods with the one-pass method being referred to again in Chapter 9. The reader is urged to refer to the references cited above for a full description of each of the methods.

The simplicity of the one-pass system is that it executes three operations during the recognition process, these are; word boundary detection, non linear time alignment and recognition. Thus recognition errors due to word boundary detection algorithm errors or due to time alignment errors are not possible (Ney, 1984). This also means that no preliminary segmentation need be performed prior to recognition, as is required for other schemes such as the two-level algorithm where beginning and endings of words must be found. The one-pass method is achieved by implementing one global time warp across the complete unknown connected phrase this is achieved by performing one time warp for each reference word template at each test frame. For example for a test sentence represented by frames i which varies from 1 to N being matched with a vocabulary of reference words represented as k which varies from 1 to K and where each reference word consists of data frames labelled j which increments from 1 to J(k), then the one-pass method calculated distance at grid points (i, j, k) for all i, j, and k. A generalized algorithm for this method would be

Initilize distances

Loop on test frames $i = 1 to N$

Loop on reference templates $k = 1 to K$

Loop on frames of reference templates $j = 1$ to $J(k)$

Compute the cumulative distance and a backpointer to the previous distance measures.

Trace back the best path from the grid point at a template ending frame with minimum total distance using the accumulated distances to find the recognized sentence.

For each word segment, only the best template, that is the template with minimised accumulated distance at its ending frame for the word segment is kept.

An illustration of this method showing the optimal path computed by the one-pass method is given in Fig. 5.8(a).

Sakoe derived the two-level algorithm, for which, on the first level, all reference patterns are systematically matched, using a band method of DTW, at all word positions of the test pattern. On the second level the algorithm sorts all the distance scores generated and the optimal estimate of the unknown sequence of words is obtained by minimising the total distance of all possible word sequences. One disadvantage of this method is that segmentation is required to find endpoints of words within phrases. Another disadvantage of this method is that it does have a large computational overhead. An illustration of this method is given in Fig. 5.8(b).

The level-building method, derived by Myers and Rabiner(1980) makes use of the property that the matching of all possible word sequences can be performed by successive concatenation of reference patterns as long as the length (in words) of the test

pattern is previously known. For the level-building method, DTW is used to match the test string against all possible sequences of reference templates (to a particular length) and finds a set of reference templates which produce a global minimum distance over all the reference words tested. The algorithm has the shortcomings that the maximum number of levels L must first be specified which may be known apriori (such as with telephone numbers) or must be calculated prior to recognition (by some form of word beginning/ending calculation). The method operates by finding the best dynamic path (that is finding the best path for each reference pattern) finishing at a point m_k for the kth reference word where m_k must lie between the constraints set by the DTW algorithm which is also dependent on the calculation level (refer Fig. 5.8(c)) so that for level l

$$m_1(l) \leq m_k \leq m_2(l). \tag{5.17}$$

By computing all possible distances within the DTW constraint this method is able to retain only those paths with the best accumulated distances for each ending point m_k at each level l. Also note that multiple distance paths for a reference templates can be calculated so that all optimal word distance within the constraints of (5.17) can be found. Finally the word string is found by backtracking from the final optimum distance through to the first.

Because an exhaustive search over all reference word to find this global minimum is not trivial many approaches for solving this problem in an efficient manner have been explored. One method is to only save the best string or strings at each level while those with poor distance scores are discarded hence reducing computation and producing an algorithm like the two-level method. Fig. 5.8(c) illustrates this method of connected word recognition (Rabiner and Levinson, 1981) which could be represented by the general algorithm:

Reported differences in recognition accuracies between these three systems have been insignificant (Godin and Lockwood, 1989) but there are major differences in the complexity and computational cost of each algorithm. The two-level algorithm is highly complicated and is not a left-to-right process, making it inapplicable for real-time applications. The level-building algorithm is more efficient however it is not a left-to-right procedure either. The one-pass scheme requires fewest computations and it has a left right flow making it applicable for real-time applications and is considered further in Chapter 9.

5.1.6 Reduction of Computational Burden

DTW is a very reliable method of time alignment between a reference and test pattern (Brown and Rabiner, 1982). However one of its major disadvantages is the heavy computational burden required to find the optimal time alignment path. Several alternative procedures have been proposed for reducing the computational burden of DTW algorithms. These methods can be categorised as:

- Reducing the overall number of local distance computations of the DTW (Bisiani and Waibel, 1982) by terminating the calculations of a reference word template when the local distance exceeds a threshold. One way this is achieved is by setting a constant threshold on the accumulated calculated distance. Other threshold methods, such as the branch and bound(BB) and the beam search(BEAM) use dynamic threshold techniques changing threshold limits based on distance calculation. These dynamic threshold methods require that the warping paths to all reference templates be computed simultaneously requiring large amounts of memory but allowing the distance threshold to be continually updated while the reference templates are being matched. A further method of reducing the number

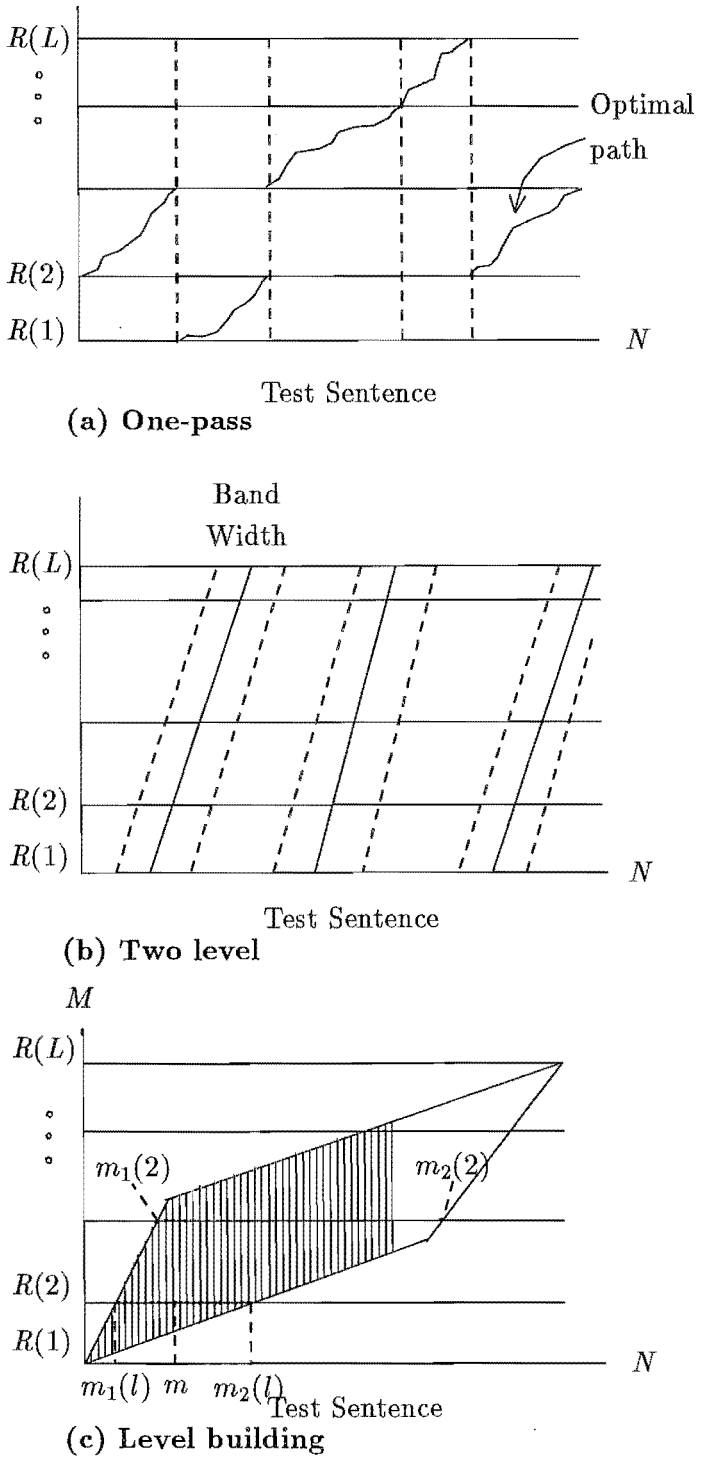


Figure 5.8. Illustration of DTW used for connected word recognition. The reference words are denoted by $R_1, R_2..R_L$ for an L word vocabulary and the test sentence is denoted as $T_1, T_2..T_M$ (a) The one-pass method compares the continuous test sentence to individual reference words. (b) The two-level method requires the test sentence to be segmented into its individual words $T_1, T_2..T_M$ and each word compared to the reference words. In the figure the number of words in the sentence, M, is shown as 4. Word comparison is via a band DTW method, which is shown here as dotted lines on each side of the warping path. (c) The level-building method compares every reference word along the total length of the test sentence, building a matched reference string at each level until the procedure solves to find the global optimum combination of reference words of a particular length.

of local distance computations is by calculating only every second or third distance value (Furui, 1986) so that the warping path is calculated only with these, fewer, local distance calculations.

- Reducing the computational cost of each DTW by applying computational windows or other related techniques to constrain the warping path. The warping path is often constrained by applying efficient search techniques such as the *ordered graph search technique* (Brown and Rabiner, 1982). These methods reduce the number of distance calculations by finding the best area for the warping function to search. Complicated methods of optimally defining the best path so that the constraint on the warping path does not reduce accuracy can greatly increase the control structure of the calculations.

- Reducing the required number of DTW procedures for a vocabulary, usually by using some initial rough classification limiting the dictionary of reference words to be searched (Glassman, 1985; Rabiner *et al.*, 1982). One way this is achieved is by some initial calculation which reduces the vocabulary to a small set of reference word templates. The initial classification can be based on such information as the test word's phonemes pattern or voiced/unvoiced pattern. Another method is to use a set of distance calculations, calculated initially from the reference templates, to compare with the test word (Vidal and Lloret, 1988; Vidal *et al.*, 1988). The test word need only be compared, using a DTW calculation, with a few of the reference words in the set to ascertain its relative position to all other reference words, based on this set of initial distance calculations.

Many of these methods have limited success, with recognition accuracies falling when computations are reduced. Typically calculation savings of between 50-70% are quoted, with a corresponding 0-10% increase in errors.

5.2 HIDDEN MARKOV MODELS (HMM)

Baker(1974) was first to propose hidden Markov modelling for word recognition. The HMM recognition process is the basis of his sophisticated connected speech recognition system known as the DRAGON speech understanding system. The following discusses the HMM method in a general sense. Most of the following text has been taken from (Rabiner, 1989).

The hidden Markov model is derived from the more general Markov modelling where both methods represent a process with a finite set of states and probabilities of moving from one state to another. Moreover a first order Markovian process is characterised by movements from one state to another being only dependent on that single past state, that is the probability of passing from one state at instant $t-1$ to the next state at instant t depends only on the state at $t-1$ and is independent of all earlier states. For a hidden Markov model however the states cannot be observed directly and are *hidden*, instead some stochastic process is observed which is assumed to have been caused by the set of states. Thus, representing a word with a hidden Markov model therefore consists of a number of states where the states produce some process showing information of the word with respect to time. Each state of the Markovian process possess distinctive properties which can be represented by some measurable quantity, however the relationship between the states and some definable speech property is difficult to quantify even though the states have been related to the phonetics of a word (Nag *et al.*, 1986).

For a Markovian process there are two probability matrices representing the initial state and the movement of the model from state to state, the initial probability and the

transitional probability. The initial probability is the probability of a HMM word model representing a word with a particular initial state, and the transitional probability is the probability of the word model moving from one state to another state when representing a word. The HMM method resolves the problem of time normalisation by the duration in a particular state.

The initial probability, or the probability of the process beginning at any particular state, is represented by $u_1, u_2...$ for a first order K-state Markov chain. The initial probabilities are written in vector form as,

$$\mathbf{u^t} = [u_1, u_2, ...u_K], \tag{5.18}$$

where $\mathbf{u^t}$ represents the transpose of the initial probability vector, and $\sum_{j=1}^{K} u_j = 1$ where $u_j \geq 0$ for all j. A transitional probability matrix, or the probability of the process moving from one state to any other state, can be written such that

$$\mathbf{V} = [v_{ij}] \quad i,j = 1..K, \tag{5.19}$$

where $\sum_{j=1}^{K} v_{ij} = 1$ for all i, which is the probability of making a transition from state i to j given the current state is state i.

For a Markov model any word can be written as a sequence (or collection) of information states represented as,

$$\theta = \theta(k)_0\theta(k)_1\theta(k)_2...\theta(k)_T, \tag{5.20}$$

where each θ at time t, $\theta(k)_t$, can be chosen from a set of K states for a K state Markov chain, and in this instance is chosen as state $\theta(k)_t$, the state at time t. For a coin tossing example the sequence of information would be a sequence of heads and tails, while the number of states, K, would equal 2. For ten coin tosses the Markov model sequence may be written as

$$\begin{aligned} \theta &= \theta(1)_0\theta(2)_1\theta(2)_2\theta(1)_3...\theta(2)_9 \\ \theta &= \text{head tail tail head head head tail head tail tail} \end{aligned} \tag{5.21}$$

where $\theta(1)$ represents state $k = 1$ (head) and $\theta(2)$ represents state $k = 2$ (tail).

For a hidden Markov model (HMM) the states $\theta = \theta_0\theta_1...\theta_T$ cannot be observed directly, but instead a stochastic process $\mathbf{S} = S_1S_2...S_T$ is observed. For speech this stochastic process may be sets of calculable features such as LPCs (refer §4.3). After each state transition an observation output $\mathbf{S}$ is produced according to a probability distribution which only depends on the current state. For a K state model there are K such observation probability distributions.

A complete specification of an HMM requires specification of the number of states in the model (K), the number of distinct observation symbols per state, N (for a coin tossing experiment, such as the one discussed above, K may equal the output from a coin or the number of coins being tossed (where the coins may be biased) and be denoted as k_1, k_2, while N equates to the physical output of the system being modelled, that is 2 representing either head or tail), specification of observation symbols, and the specification of the three probability measures $\mathbf{V}$, $\mathbf{u}$, and $\mathbf{F}$, where $\mathbf{F}$ is the set of observation symbol probability distributions at each state θ_i, that is, the set of $f_{\theta_i}(n)$, where

$$\begin{aligned} f_{\theta_i}(n) &= P[k_n \text{ at } t|\theta_t = \theta_j] \quad 1 \leq j \leq K \\ &\qquad\qquad\qquad\qquad\qquad\quad 1 \leq n \leq N \end{aligned} \tag{5.22}$$

For convenience, the compact notation

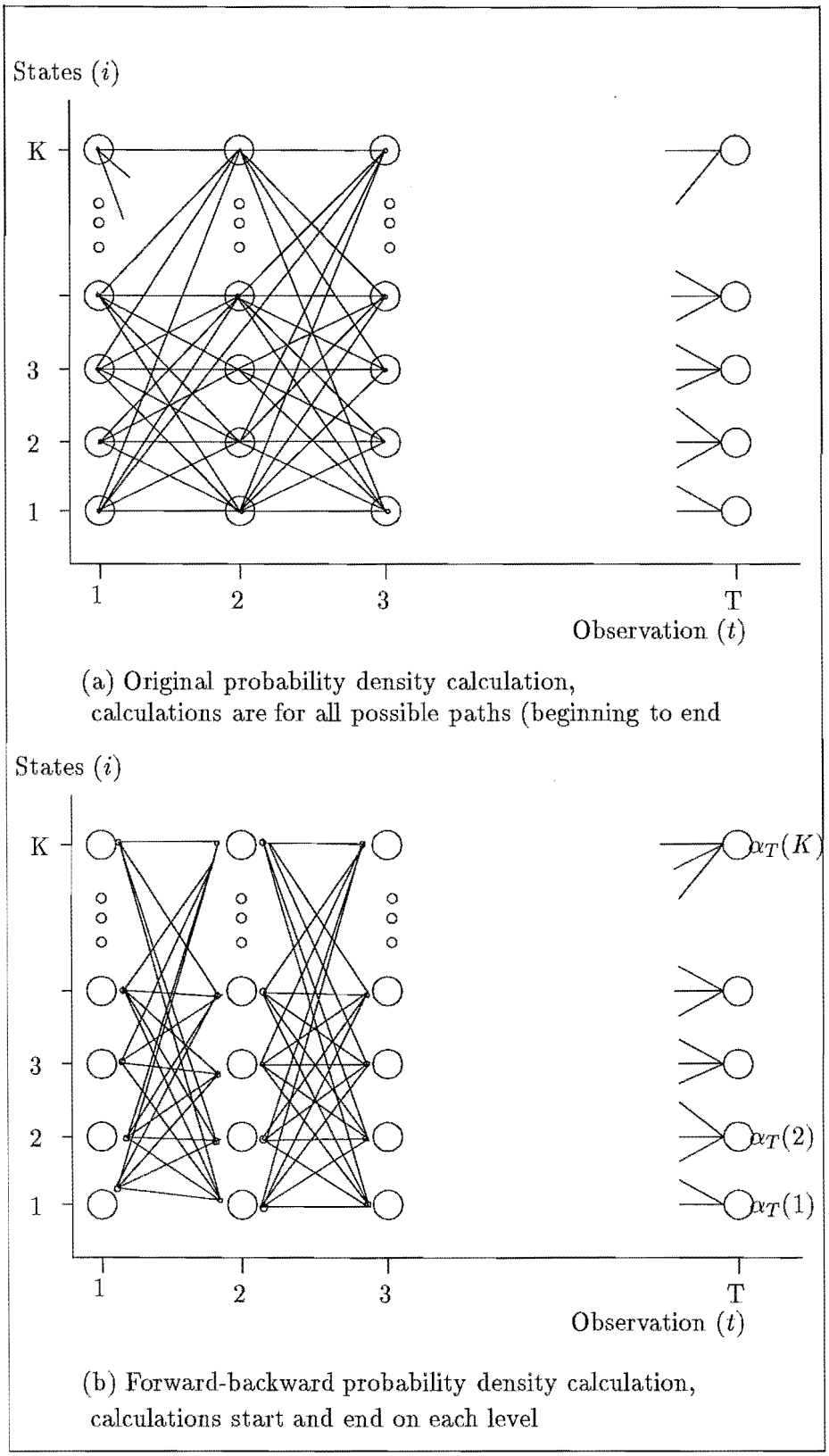


Figure 5.9. Illustration of the methods used to evaluate the probability densities for a Markov model. (a) Original probability density calculation showing calculations over all possible paths. This method has exponential calculation growth with observations and states. (b) Forward-backward probability density calculation, calculations begin and end at each observation level. This method has linear calculation growth with observations and states.

$$\mathbf{M} = (\mathbf{V}, \mathbf{u}, \mathbf{F}) \tag{5.23}$$

is used to indicate the complete parameter set of the model.

For the form of HMM discussed above there are three basic problems of interest. These three problems must be solved for the model to be useful in real-world applications. The problems are the following:

Problem 1: Given the observation sequence $\mathbf{S} = \mathbf{S_1 S_2 ... S_T}$, and a model $\mathbf{M} = (\mathbf{V}, \mathbf{u}, \mathbf{F})$ how is $P(S|M)$ computed, that is the probability of an observation set given a model M?

Problem 2: Given the observation sequence $\mathbf{S} = \mathbf{S_1 S_2 ... S_T}$, and the model $\mathbf{M}$, how is a corresponding state sequence $\theta = \theta_1 \theta_2 ... \theta_T$ chosen, which is optimal in some meaningful sense (ie best *explains* the observations)?

Problem 3: How are the model parameter $\mathbf{M} = (\mathbf{V}, \mathbf{u}, \mathbf{F})$ adjusted to maximise $P(S|M)$?

Problem 1 is the evaluation problem, namely given a model and a sequence of observations, how do we compute the probability that the observed sequence was produced by the model. This problem can also be viewed as one of scoring how well a given model matches a given observation sequence. The latter viewpoint is particularly useful. For example, if the case is considered in which a decision is to be made about several competing models, the solution to Problem 1 allows the model to be chosen which best matches the observations.

Problem 2 is the one which uncovers the hidden part of the model, ie, to find the *correct* state sequence. However, for all but the case of degenerate models, there is no *correct* state sequence to be found. Hence for practical situations, an optimality criterion to solve this problem is used to find the best possible.

Problem 3 is the one in which an attempt is made to optimize the model parameters so as to best describe how a given observation sequence comes about. The observation sequence used to adjust the model parameters is called a training sequence since it is used to *train* the HMM. The training problem is the crucial one for most applications of HMMs, since it allows the model parameters to optimally adapt to the observed training sequence.

5.2.1 Solutions to the Three basic problems of HMM

5.2.1.1 Solution to Problem 1

Problem 1 is one of calculating the probability of the observation sequence

$$\mathbf{S} = \mathbf{S_1 S_2 ... S_T}, \tag{5.24}$$

given the model $\mathbf{M}$, that is $P(S|M)$. The most straightforward way of doing this is through enumerating every possible state sequence of length T (the number of observations). Consider one such fixed state sequence

$$\theta = \theta_1 \theta_2 ... \theta_T \tag{5.25}$$

where θ_1 is the initial state. The probability of the observation sequence $\mathbf{S}$ for the state sequence 5.25 is

$$P(S|\theta, \mathbf{M}) = \prod_{t=1}^{T} \mathbf{P}(\mathbf{S_t}|\theta_t, \mathbf{M}) \tag{5.26}$$

where statistical independence of observations is assumed. Therefore it can be written

$$P(S|\theta,\mathbf{M}) = \mathbf{f}_{\theta_1}(\mathbf{S_1})\mathbf{f}_{\theta_2}(\mathbf{S_2})...\mathbf{f}_{\theta_T}(\mathbf{S_T}) \tag{5.27}$$

The probability of such a state sequence θ can be written as

$$P(\theta|\mathbf{M}) = \mathbf{u}_{\theta_1}\mathbf{v}_{\theta_1\theta_2}\mathbf{v}_{\theta_2\theta_3}...\mathbf{v}_{\theta_{T-1}\theta_T}. \tag{5.28}$$

The joint probability of $\mathbf{S}$ and θ, that is the probability that $\mathbf{S}$ and θ occur simultaneously is simply the product of the above two terms, ie

$$P(S,\theta|\mathbf{M}) = \mathbf{P}(\mathbf{S}|\theta,\mathbf{M})\mathbf{P}(\theta,\mathbf{M}) \tag{5.29}$$

The probability of S given the model is obtained by summing this joint probability over all possible state sequences θ giving

$$\begin{aligned} P(S|M) &= \textstyle\sum_{all\theta} P(S|\theta,\mathbf{M})\mathbf{P}(\theta,\mathbf{M}) \\ &= \textstyle\sum_{\theta_1,\theta_2...,\theta_T} u_{\theta_1} f_{\theta_1}(S_1) v_{\theta_1\theta_2} f_{\theta_2}(S_2) \\ &\quad ...u_{\theta_{T-1}\theta_T} f_{\theta_T}(S_T) \end{aligned} \tag{5.30}$$

The interpretation of the computation in the above equation is the following. Initially (at time t=1) the model is in state θ_1 with probability u_{θ_1}, and will generate the symbol S_1 (in this state) with probability $f_{\theta_1}(S_1)$. The clock changes from time t to $t+1$ $(t=2)$ and a transition is made to state θ_2 from state θ_1 with probability $v_{\theta_1\theta_2}$, and generate symbol S_2 with probability $f_{\theta 2}(S_2)$. This process continues in the same manner until the last transition is made (at time T) from state θ_{T-1} to state θ_T with probability $v_{\theta_{T-1}\theta_T}$ and generates symbol S_T with probability $f_{\theta_T}(S_T)$. The calculation of $P(S|M)$ by the direct definition 5.30 involves in the order of $2T.K^T$ calculations, since at every $t = 1,2,...,T$, there are K possible states which can be reached (therefore there are K^T possible state sequences), and for each such state sequence about $2T$ calculations are required for each term in the sum of 5.30. This calculation is computational unfeasible, even for small values of K and T. A more efficient procedure to solve Problem 1 is the *forward-backward* procedure(Baum and Egon, 1967)(Baum and Sell, 1968) For the forward-backward procedure, a forward variable $\alpha_t(i)$ can be defined such that

$$\alpha_t(i) = P(S_1S_2...S_t, \theta_t = \theta_i|M) \tag{5.31}$$

that is the probability of the partial observation sequence, $S_1S_2...S_t$, until time t and state θ_i at time t, given the model M. An inductive solution for $\alpha_t(i)$ can be found as follows:

1 Initialisation:

$$\alpha_1(i) = u_i f_i(S_1), \quad 1 \le i \le K. \tag{5.32}$$

2 Induction:

$$\begin{aligned} \alpha_{t+1}(j) &= \left[\textstyle\sum_{i=1}^{K} \alpha_t(i)v_{ij}\right] f_j(S_{t+1}), \quad 1 \le t \le T-1 \\ &\qquad\qquad\qquad\qquad\qquad\qquad 1 \le j \le K. \end{aligned} \tag{5.33}$$

3 Termination:

$$P(S|M) = \sum_{i=1}^{K} \alpha_T(i). \tag{5.34}$$

Step 1 initialises the forward probabilities as the joint probability of state θ_i and initial observation S_1. The induction step is the heart of the forward calculation. The computation for the induction in step 2 is performed for all states j, $1 \leq j \leq K$ for a given t; the computation is then iterated for $t = 1, 2, ..., T-1$. Finally step 3 gives the desired calculation of $p(S|M)$ as the sum of the terminal forward variables $\alpha_t(i)$ this is the case since, by definition,

$$\alpha_T(i) = P(S_1 S_2 ... S_T, \theta_T = \theta_i | M) \tag{5.35}$$

and hence $P(S|M)$ is just the sum of the $\alpha_T(i)$'s.

The computation involved in the calculation of $\alpha_t(j), 1 \leq t \leq T, 1 \leq j \leq K$ is of the order of K^2T calculations, rather then $2TK^T$ as required for the direct calculation.

The forward probability calculation is based upon the lattice (or trellis) structure shown in 5.9.

As shown in the lattice calculation figure 5.9 at time $t = 1$ only values for $\alpha_1(i), 1 \leq i \leq K$ need be calculated. At times $t = 2, 3, ..., T$ values of $\alpha_t(j)$ at $1 \leq j \leq K$ need be calculated where each calculation involves only K previous values of $\alpha_{t-1}(i)$ because each of the K grid points is reached from the same K grid points at the previous time slot.

In a similar manner backward variables $\beta_t(i)$ can also be defined such that

$$\beta_t(i) = P(S_{t+1} S_{t+2} .. S_T | \theta_t = \theta_i, M) \tag{5.36}$$

that is, the probability of the partial observation sequence for $t+1$ to the end, given state θ_i at time t and the model M. Again, $\beta_t(i)$ can be solved for inductively as

1 Initialisation:

$$\beta_T(i) = 1, 1 \leq i \leq K. \tag{5.37}$$

2 Induction:

$$\begin{aligned} \beta_t(i) &= \sum_{j=1}^{K} v_{ij} f_j(S_{t+1}) \beta_{t+1}(j), \\ & t = T-1, T-2, ..., 1, 1 \leq i \leq K. \end{aligned} \tag{5.38}$$

The initialization step 1 arbitrarily defines $\beta_T(i)$ to be 1 for all i. Step 2 says that in order to have been in state θ_i at time t, and to account for the observation sequence from time $t+1$ on, then all possible states θ_j at time $t+1$ must be considered, accounting for the transition from θ_i to θ_j (the v_{ij} term) as well as the observation S_{t+1} in state j (the $f_j(S_{t+1})$ term), and then account for the remaining partial observation sequence from state j (the $\beta_{t+1}(j)$ term). Both the forward and backward calculations are used to solve fundamental Problems 2 and 3 of HMMs.

The computation of $\beta_t(j), 1 \leq t \leq T, 1 \leq i \leq K$, requires in the order of K^2T calculations, and can be computed in a lattice structure similar to that of the forward computation and shown in figure 5.9.

5.2.1.2 Solution to problem 2

Unlike Problem 1 for which an exact solution can be given, there are several ways of solving Problem 2, namely finding the *optimal* state sequence associated with the given observation sequence. The difficulty lies with the definition of the optimal state sequence; that is there are several possible optimality criteria. For example, one possible optimality criterion is to choose the states θ_t, which are individually most likely. This optimality criterion maximises the expected number of correct individual states. To implement this solution to Problem 2, we define the variable

$$\gamma_t(i) = P(\theta_t = \theta_i | S, M) \tag{5.39}$$

that is the probability of being in state θ_i at time t, given the observation sequence S, and the model M. Equation 5.39 can be expressed simply in terms of the forward-backward variables,

$$\begin{aligned} \gamma_t(i) &= \frac{\alpha_t(i)\beta_t(i)}{P(S|M)} \\ &= \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{K}\alpha_t(i)\beta_t(i)} \end{aligned} \tag{5.40}$$

since $\alpha_t(i)$ accounts for the partial observation sequence $S_1S_2...S_t$ and state θ_i at t, while $\beta_t(i)$ accounts for the remainder of the observation sequence $S_{t+1}S_{t+2}...S_T$, given state θ_i at t. The normalization factor $P(S|M) = \sum_{i=1}^{K}\alpha_t(i),\beta_t(i)$ makes $\gamma_t(i)$ a probability measure so that

$$\sum_{i=1}^{K}\gamma_t(i) = 1. \tag{5.41}$$

Using $\gamma_t(i)$, we can solve for the individually most likely state θ_t at time t, as

$$\theta_t = argmax_{i\leq i\leq K}[\gamma_t(i)], 1 \leq t \leq T. \tag{5.42}$$

Although 5.42 maximises the expected number of correct states (by choosing the most likely state for each t), there could be some problems with the resulting state sequence. For example, when the HMM has state transitions which have zero probability ($v_{ij} = 0$ for some i and j), the *optimal* state sequence may, in fact, not even be a valid state sequence. This is due to the fact that the solution of 5.42 simply determines the most likely state at every instant, without regard to the probability of occurrence of sequences of states.

One possible solution to the above problem is to modify the optimality criterion. For example, one could solve for the state sequence that maximises the expected number of correct pairs of states $\{\theta_t, \theta_{t+1}\}$, or triples of states $\{\theta_t, \theta_{t+1}, \theta_{t+2}\}$, etc. Although these criteria might be reasonable for some applications, the most widely used criterion is to find the single best state sequence (path), ie to maximise $P(\theta|\mathbf{S},\mathbf{M})$ which is equivalent to maximising $P(\theta, \mathbf{S}|\mathbf{M})$. A formal technique for finding this single best state sequence exists, based on dynamic programming methods, and is called the Viterbi algorithm.

The Viterbi algorithm is a method to find the single best state sequence $\theta = \{\theta_1, \theta_2...\theta_q\}$, for the given observation sequence $S = \{S_1, S_2, ...S_T\}$. For such a case is is necessary to define the quantity

$$\eta_t(i) = max_{\theta_1,\theta_2,...\theta_{t-1}} P[\theta_1\theta_2...\theta_t = i, S_1S_2...S_t|M] \tag{5.43}$$

that is, $\eta_t(i)$ is the best score (highest probability) along a single path, at time t, which accounts for the first t observations and ends in state θ_j. By induction we have

$$\eta_{t+1}(j) = [\max_i \eta_t(i)v_{ij}].f_j(S_{t+1}). \tag{5.44}$$

To actually retrieve the state sequence, we need to keep track of the argument which maximised 5.44, for each t and j. This is achieved via the array $\lambda_t(j)$. The complete procedure for finding the best state sequence can now be stated as follows:

1 Initialisation

$$\begin{aligned} \eta_1(i) &= u_i f_i(S_1), \quad 1 \leq i \leq K \\ \lambda_1(i) &= 0. \end{aligned} \tag{5.45}$$

2 Recursion:

$$\begin{aligned} \eta_t(j) &= max_{1\leq i\leq K}[\eta_{t-1}(i)v_{ij}]f_i(S_t) \quad 2 \leq i \leq T \\ & \qquad\qquad\qquad\qquad\qquad\quad 1 \leq j \leq K \\ \lambda_t(j) &= argmax_{1\leq i\leq K}[\eta_{t-1}(i)v_{ij}], \quad 2 \leq t \leq T \\ & \qquad\qquad\qquad\qquad\qquad\quad 1 \leq j \leq K \end{aligned} \tag{5.46}$$

3 Termination

$$\begin{aligned} P^* &= max_{1\leq i\leq K}[\eta_T(i)] \\ \theta_T^* &= argmax_{1\leq i\leq K}[\eta_T(i)] \end{aligned} \tag{5.47}$$

4 Path (state sequence) backtracking:

$$\theta_t^* = \lambda_{t+1}(\theta_{t-1}^*), t = T-1, T-2, ..., 1. \tag{5.48}$$

It should be noted that the Viterbi algorithm is similar (except for the backtracking step) in implementation to the forward calculations of the forward-backward procedure discussed previously. The major difference is the maximisation over previous states which is used in place of the summing procedures in the forward-backward procedure. It should also be clear that a lattice (or trellis) structure efficiently implements the computation of the Viterbi procedure.

5.2.1.3 Solution to problem 3

The third, and by far the most difficult, problem of HMM is to determine a method to adjust the model parameters $(\mathbf{V}, \mathbf{u}, \mathbf{F})$ to maximise the probability of the observation sequence given the model. There is no known way to analytically solve for the model which maximises the probability of the observation sequence. In fact, given any finite observation sequence as training data, there is no optimal way of estimating the model parameters. We can, however, choose $M = (\mathbf{V}, \mathbf{u}, \mathbf{F},)$ such that $P(S|M)$ is locally maximised using an iterative procedure such as the Baum-Welch method. In this section an iterative procedure, based on the classic work of Baum and his colleagues is discussed which can be used to choose model parameters.

In order to describe the procedure for reestimation (iterative update and improvement) of HMM parameters, $\zeta_T(i,j)$, is first defined as the probability of being in state θ_i at time t, and state θ_j at time $t+1$, given the model and the observation sequence, i.e.

$$\zeta_t(i,j) = P(\theta_t = \theta_i, \theta_{t+1} = \theta_j|\mathbf{S}, \mathbf{M}). \tag{5.49}$$

The sequence of events leading to the conditions required by equation 5.49

is illustrated in 5.10. It should be clear, from the definitions of the forward and backward variables, that $\zeta_t(i,j)$ can be written in the form

$$\begin{aligned} \zeta_t(i,j) &= \frac{\alpha_t(i)v_{ij}f_i(S_{t+1})\beta_{t+1}(j)}{P(S|M)} \\ &= \frac{\alpha_t(i)v_{ij}f_j(S_{t+1})\beta_{t+1}(j)}{\sum_{i=1}^{K}\sum_{j=1}^{K}\alpha_t(i)v_{ij}f_j(\theta_{t+1})\beta_{t+1}(j)} \end{aligned} \tag{5.50}$$

where the numerator term is just $P(\theta_t = \theta_i, \theta_{t+1} = \theta_j, \mathbf{S}|\mathbf{M})$ and the division by $P(\mathbf{S}|\mathbf{M})$ gives the desired probability measure.

The definition of $\gamma_t(i)$ has been given previously as the probability of being in state θ_i at time t, given the observation sequence and the model; hence we can relate $\gamma_t(i)$ to $\zeta_t(i,j)$ by summing over j, giving

$$\gamma_t(i) = \sum_{j=1}^{K}\zeta_t(i,j). \tag{5.51}$$

If $\gamma_t(i)$ is summed over the time index t, a quantity is obtained which can be interpreted as the expected (over time) number of times that state θ_i is visited, or equivalently, the expected number of transitions made from state θ_i (if the time slot $t = T$) is excluded

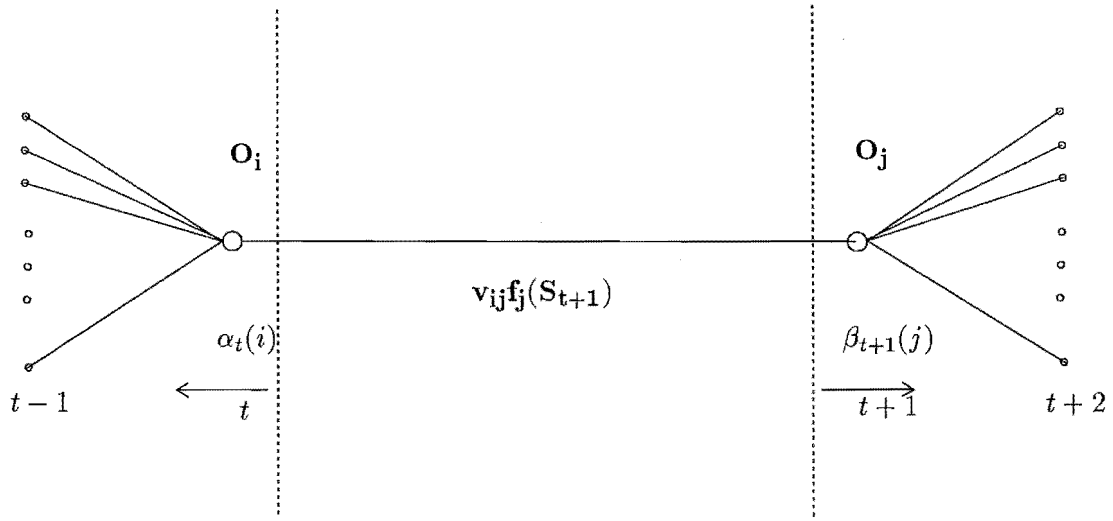


Figure 5.10. Illustration of the sequence of operations required for the computation of the joint event that the system is in state θ_i at time t and state θ_j at time $t+1$

from the summation). Similarly, summation of $\zeta_t(i,j)$ over t (from $t=1$ to $t=T-1$) can be interpreted as the expected number of transitions from state θ_i to state θ_j, that is

$$\begin{array}{lcl} \sum_{t=1}^{T-1}\gamma_t(i) & = & \text{the expected number of transitions from } \theta_i \\ \sum_{t=1}^{T-1}\zeta_t(i,j) & = & \text{the expected number of transitions from } \theta_i \text{ to } \theta_j \end{array} \tag{5.52}$$

using the above formulas and the concept of counting event occurrences, a method for reestimation of the parameters of HMM can be found. A set of reasonable reestimation formulas for M, v, and F are

$$\begin{aligned} \hat{M}_i &= \text{expected frequency (number of times) in state } \theta_j \text{ at time } (t=1) = \gamma_1(i) \\ \hat{v}_{ij} &= \frac{\text{expected number of transitions from state } \theta_i \text{ to state } \theta_j}{\text{expected number of transitions from state } \theta_i} \\ &= \frac{\sum_{t=1}^{T-1}\zeta_t(i,j)}{\sum_{t=1}^{T-1}\gamma_t(i)} \\ \hat{f}_{j(n)} &= \frac{\text{expected number of times in state } j \text{ and observing symbol } k_n}{\text{expected number of times in state } j} \\ &= \frac{\sum_{t=1}^{T}[S_t=c_n]\gamma_t(j)}{\sum_{t=1}^{T}\gamma_t(j)} \end{aligned} \tag{5.53}$$

The current model can be defined as $\mathbf{M}=(\mathbf{V},\mathbf{u},\mathbf{F})$ which can be used to compute the right-hand sides of equations 5.53. The reestimation model can be defined as

$$\hat{\mathbf{M}}=(\hat{\mathbf{V}},\hat{\mathbf{u}},\hat{\mathbf{F}}), \tag{5.54}$$

determined from the left-hand sides of equations 5.53. It has be proven by Baum and his colleagues (Baum and Sell, 1968) that either; the initial model $\mathbf{M}$ defines a critical

point of the likelihood function in which case $\hat{\mathbf{M}} = \mathbf{M}$; or model $\hat{\mathbf{M}}$ is more likely than model $\mathbf{M}$ in the sense that $P(S|\hat{\mathbf{M}}) > P(\mathbf{S}|\mathbf{M})$, that is a new model $\hat{\mathbf{M}}$ has been found from which the observation sequence is more likely to have been produced.

Based on the above procedure, if $\hat{\mathbf{M}}$ is used iteratively in place of $\mathbf{M}$ and the reestimation calculations are repeated, the probability of $\mathbf{S}$ being observed from the model can be improved to a limit. The final result of this reestimation procedure is called a maximum likelihood estimate of the HMM. It should be pointed out that the forward-backward algorithm leads to local maxima only, and that in most problems of interest, the optimisation surface is very complex and has many local maxima.

5.2.2 Structure of the HMM

Structures used for Markov modelling give the allowable movement from one state to another. The structures fall into three major categories; unconstrained (the ergodic model), constrained serial (the left to right model) and constrained parallel (the parallel left to right model) (Rabiner *et al.*, 1986) as shown in Fig. 5.11.

In the unconstrained, or ergodic, model (Fig. 5.11(a)) a transition from any state to any other state can be made, that is all the transitional probabilities v_{ij}'s are allowed to be nonzero.

The constrained serial (Fig. 5.11(b)) and the constrained parallel models (Fig. 5.11(c)) are special cases of Markov chains used for isolated word recognition which give temporal structure because they move from left to right. They also have the properties that:

- the first observation is produced while the Markov chain is in a distinct state called the starting state, - q_1.
- The last observation is generated while the Markov chain is in a distinct state called the final or absorbing state - q_N.
- Once the Markov chain leaves a state, that state cannot be revisited at a later time.

The left-to-right nature of these models limits the state transitions matrix $\mathbf{V}$ to an upper triangular matrix. Movement along the paths is dependent on the serial/parallel model constraint. The serial construct generally constrains the model to proceed sequentially through the states, although individual states can be skipped, whereas the parallel model allows for multiple paths.

Each of the model structures of Fig. 5.11 can be generalised to include an arbitrary number of states. However their appears to be no way of choosing the optimum number of states needed for a word model, as they are not physically related to any single observable phenomenon.

5.2.3 Applications of HMMs

In the speech field HMMs were initially used for isolated word recognition, however their use in speech has since widened. Primarily used for large vocabulary word recognition with the advent of systems such as Dragon (Baker, 1974), Tangora (Averbuch *et al.*, 1986) and SPHINX (Lee, 1988), the Markov modelling process is a powerful method of modelling any quasi-stationary signal. Ljolje and Fallside(1987) discussed a method where hidden Markov models were used for the modelling and recognition of prosodic patterns. The prosody of the speech was measured in fundamental frequency, timing and intensity of the sounds. Recognition rates of around 93% were obtained demonstrating the potential of HMMs for modelling any speech patterns.

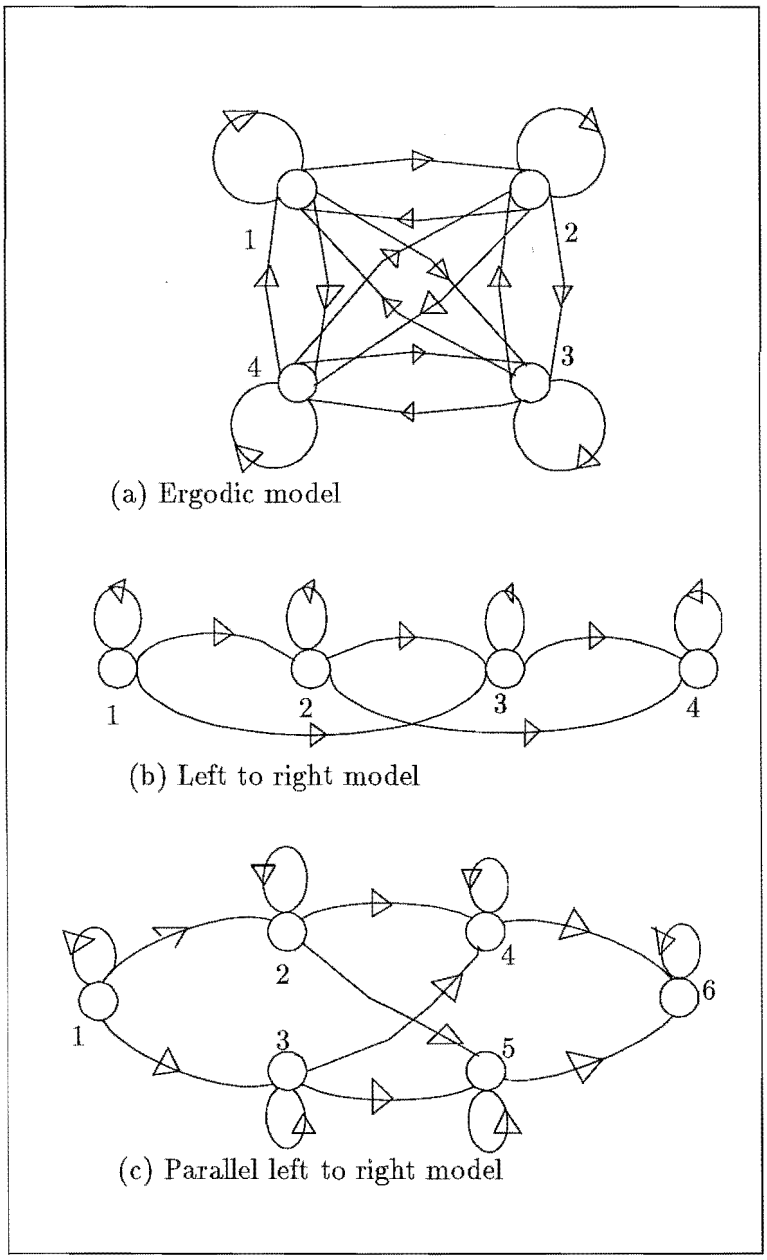


Figure 5.11. Illustration of a set of general Markov model structures. (a)Ergodic model allows any states to move to any other state. (b) Left-to-right model, movement is only allowed to a forward state, adding time sequence to the model. This model is usually used for speech. (c) Parallel left-to-right model, movement is only allowed to a forward state adding time sequence to the model, however this model has a more complicated structure than (b) not necessarily equivalent to speech transitions.

5.3 DYNAMIC TIME WARPING AND HIDDEN MARKOV MODELLING

A relationship exists between DTW with LPC distance measures (refer §6.4) and HMM with Gaussian autoregressive densities as discussed by Juang(1984). This relationship was discussed for limited HMM models where the probability of transiting from one state to another or of being in a particular state are equal. Initially by considering the dynamic programming representations of two speech sequences, one as a reference and

one as a test, ie.,

$$W = w_1, w_2 ... w_{Tw}, \tag{5.55}$$

$$Y = y_1, y_2 ... y_{Ty}, \tag{5.56}$$

a warping function is said to exist such that a correspondence can be calculated between these two sequences,

$$\begin{aligned} t_y &= \phi(t_w), \quad t_w = 1, 2 ... T_w, \\ t_w &= \zeta(t_y), \quad t_y = 1, 2 ... T_y. \end{aligned} \tag{5.57}$$

The warping function is found by minimising the distance between the two function, either,

$$D_\phi = \sum_{t_w=1}^{T_w} d[w_{t_w}; y_{\phi(t_w)}], \tag{5.58}$$

or

$$D_\zeta = \sum_{t_y=1}^{T_y} d[y_{t_y}; w_{\zeta(t_y)}]. \tag{5.59}$$

For the HMM a T_w state model $\mathbf{M_W} = (\mathbf{V_W}, \mathbf{u_W}, \mathbf{F_W})$ can be written for the reference pattern with transitional probabilities V_w, a T_w by T_w matrix with probabilities $v_{ij} = 1/T_w$ for all $i, j = 1..T_w$ (all transitional probabilities equal) and an initial probability u_w with $u_i = 1/T_w$ for any $i = 1..T_w$. Similar equations can be written for a T_y state test pattern. Each state has a maximum likelihood of occurring based on an observation w_i represented by its Gaussian autoregressive density f_i for $i = 1$ to T_w. Each f_i is defined by the parameter pair (a_w, σ_w^2), which is its set of LPC values and gain values. The log likelihood that the function is word W for state θ_w given the model M_w, is given by the maximum log likelihood,

$$\begin{aligned} log[f(W, \theta_w | M_w)] &= log[(1/T_w)^{T_w+1} \textstyle\prod_{i=1}^{T_w} f_i(w_i | a_{w,i}, \sigma_{w,i}^2)] \\ &= -\tfrac{N}{2}[\textstyle\sum_{i=1}^{T_w} log(2\pi\sigma_{w,i}^2) + T_w] - (T_w + 1) log T_w \\ &= max\theta \in \{\theta\}_{T_w}\{log f(W, \theta | M_w)\}, \end{aligned} \tag{5.60}$$

where $\{\theta\}_{(T_w)}$ denotes the set of all state sequences with length T_w. A similar model can be defined for $M_y = (V_y, u_y, F_y)$.

Correspondence between the two sequences can be written by ζ-warping sequence θ_y giving the likelihood measure. The log difference between a maximum likelihood and log likelihood sequence is hence,

$$log f(Y, \theta_y | M_y) - log f(Y, \zeta(\theta_y) | M_w) = N/2 D_\zeta. \tag{5.61}$$

The accumulative distortion D_ζ in dynamic time warping is directly related to the likelihood difference between the two models. To express the density of obtaining pattern Y from model M_w or $f(Y|M_w)$ is,

$$\begin{aligned} f(Y|M_w) &= \textstyle\sum_{all\zeta} f(Y, \zeta(\theta_y) | M_w), \\ &= f(Y, \theta_y | M_y) \textstyle\sum_{all\zeta} exp(-N/2(D_\zeta)). \end{aligned} \tag{5.62}$$

These equations demonstrate that determining a warping function ζ in dynamic time warping is equivalent to finding the state sequence that maximises the density $f(Y, \zeta(\theta_y)|M_w)$.

This model holds no time constraints on the warping function and hence forward and backward warping functions are allowed. As pointed out by Juang(1984) a reversed 'we' may sound very close to a 'you' and so transition constraints are important.

Changing the transitions in the HMM maps to particular constraints imposed in the DTW algorithm and comparisons can be made. In such a case, where the underlying transition structure is equiprobable, dynamic time warping is equivalent to the probabilistic modelling technique.

Russell *et al*(1983) also discuss and compare DTW and HMM. They point out that these two recognition algorithms have many similarities stating that classification by DTW using Euclidean distance is precisely equivalent to Viterbi recognition using HMM with Gaussian states and unit covariance matrices. However the two methods differ significantly in their approaches to word-model generation. The obvious difference is that the DTW scheme simply involves storing one or more examples of the word patterns for each vocabulary word while HMM must estimate parameters for each word. Hence the major problem with HMM is to obtain enough representations of a word to produce an acceptable model.

Combining both methods Russell *et al*(1983) introduce a locally constrained dynamic time-warping model where the time-normalisation process is constrained using a first-order Markov model. The cumulative distance between an unknown word-pattern and a reference template is no longer simply a sum of distances between aligned vectors, it also includes penalties which are derived from probabilities and used for time-scale distortion at each point on the time-registration path. This approach thereby effectively combined the two techniques into the same recogniser.

Other tests of equivalence between the two methods have been to compare their recognition accuracies. Rabiner *et al*(1983) compared results on HMM and DTW using LPC based recognition. In these tests the DTW performed, on average, 2% better then the HMM. The lower accuracy for the HMM system may have been due to the limited training set. Svendsen *et al*(1989) also noted lower accuracy for the HMM system when testing whole word DTW and HMM methods. Recognition scores of 97.3% for the DTW and 90.0% for the HMM models were achieved. This therefore points to DTW being better suited to a system that can use only limited training data.

5.4 CHOOSING A RECOGNITION SYSTEM

Each of the HMM and DTW methods has advantages and disadvantages. The HMM method, although having lower computation for recognition requires a large training set to perform adequately. Training HMMs can be time consuming and requires large amounts of memory. However, once trained, the HMM method operates fast and requires little memory space. Thus the HMM method is suited for a system which does not require re-training such as a speaker independent operating recogniser.

The advantageous of the DTW method is that it can run effectively from limited training data. Re-training a DTW system is simply a matter of saving the word templates of each new speaker and although this is easily achieved this method of 'training' does require large amounts of memory for storage. Although DTW takes longer then HMM to recognize a word it can be easily programmed to operate in real-time with specialised DSP chips and limited vocabulary (below 100 reference words). Methods of reducing computation and improving the speed of operation, as discussed in §5.1.6, are well known and easy to implement with little or no accuracy reduction. A DTW system is, therefore, more suitable to a real-time base multi-user but speaker-dependent system that will require retraining for each new speaker. For the recognition experiments discussed in Chapters 7 and 8 it was decided to use a DTW scheme.

Chapter 6

INVESTIGATION OF RECOGNITION DISTANCE MEASURES

In this chapter some of the distance measures that are commonly used for word recognition and introduced in §5.1.3, are explored. Important points are emphasised through presentation of practical examples. A thorough examination and evaluation is given on a set of distance measures generally used in speech and word recognition systems. The distance measures discussed fall into three main classes:-

1 euclidean distance measures (refer §6.2), which include cepstral distance measures such as linear quefrency weighted, root power sum (RPS) quefrency weighted, liftered cepstral, and the Mahalanobis measures,

2 directional cosine distance measures, which include the cepstral projection (angle) measure (refer §6.3),

3 probability based distance measures (refer §6.4), which include the likelihood, the log likelihood and the Itakura-Saito measures.

6.1 PROPERTIES OF DISTANCE MEASURES

Gray and Markel(1976) identified four basic properties which the distance measure, $d(x, y)$, must satisfy to be a useful indicator of the lack of similarity between two frames of speech x and y. These are:-

1 symmetry $d(x, y) = d(y, x)$,

2 positive definiteness

$$\begin{aligned} &d(x, y) > 0 \quad \text{for } x \neq y \\ &d(x, x) = 0, \end{aligned} \tag{6.1}$$

3 the distance measure should have a physically meaningful interpretation in the frequency domain. (The frequency content of speech is considered to have the most perceptual relevance.),

4 it should be possible to efficiently evaluate $d(x, y)$.

If the measure is to be regarded as a distance *metric* a fifth property is also considered necessary. This fifth property, known as the *triangle inequality*, requires the distance measure to satisfy the constraint that,

$$d(x, y) \leq d(x, z) + d(y, z). \tag{6.2}$$

All of the distance measures discussed here satisfy at least four of the five properties discussed above. However, they all have distinct characteristics and these are discussed in detail in the following sections for each of the distance measures.

6.2 EUCLIDEAN DISTANCE

A Euclidean distance gives the shortest distance between two vectors in Euclidean space. For example, the Euclidean distance between two feature vectors F_1 and F_2 in a two dimensional space, whose positions are given by $(f_1(1), f_1(2))$ and $(f_2(1), f_2(2))$, can be calculated as,

$$\begin{aligned} d^2_{Euclid}(F_1, F_2) &= w(1)(f_1(1) - f_2(1))^2 + w(2)(f_1(2) - f_2(2))^2, \\ &= \textstyle\sum_{i=1}^{2} w(i)(f_1(i) - f_2(i))^2, \end{aligned} \tag{6.3}$$

where each component of the distance measure is weighted by a constant $w(i)$. The standard Euclidean distance weighting $w(i)$ is unity for all components. However, variations of the Euclidean measure have been defined which have different weightings.

The Euclidean distance can be extended to N dimensions by the general formula,

$$d^2_{Euclid}(F_1, F_2) = \sum_{i=1}^{N} w(i)(f_1(i) - f_2(i))^2. \tag{6.4}$$

The Euclidean distance satisfies all five of the distance measure properties discussed above and because of this the Euclidean measure is a distance metric. The advantage of the Euclidean distance metric is that it is one of the simplest and most easily calculated distance metrics. The standard Euclidean distance, with all weighting factors unity, has the ability to find meaningful distances between vectors that have components which span an equivalent magnitude range and so have equal significance.

For features with components of differing significance or differing magnitude scales, non-equally weighted Euclidean distance measures can be used. One such measure, known as the *Mahalanobis* metric, weights the feature components by the inverse of the covariance matrix, V, where the covariance matrix specifies the amount of correlation between two vector components. For speech (and particularly speech recognition) the covariance matrix used is the *intra-word* or *within-word* covariance matrix, V_{wi}, which is the word variance between one of the reference vectors for a word (usually the mean word vector) and all other vector representations for that word. The variance is calculated over the whole word by employing some time normalising procedure such as DTW (refer Chapter 5).

The Mahalanobis distance between feature vectors F_1 and F_2 is,

$$d_{Mal}(F_1, F_2) = (F_1 - F_2)V_{wi}^{-1}(F_1 - F_2)^T. \tag{6.5}$$

To determine the usefulness of the Mahalanobis distance measure consider the problem of classifying an individual unknown, x_0 a vector set size p of samples, into one of either two known categories (with common covariances). Say π_0, π_1, π_2 denote the three populations, where π_0 is the population of the unknown. Each population is representable by a multivariate normal distribution $N(\hat{\pi}, V)$ where $\hat{\pi}$ represents the distribution's mean and V represents the distribution's covariance. In general, for any multivariate normal distribution let X be an observation matrix of independent samples of x a p-component vector where $x \simeq N(\hat{x}, V)$ such that $\hat{x}$ is the mean of the distribution and V is the distribution's covariance and V is positive definite. Let $x_1, x_2, ..., x_N$ be independent samples on x. The observation matrix of x can be written as X such that

$$X \equiv (x_1, x_2, ..., x_N) \equiv \begin{bmatrix} x_{11} & x_{12} & ... & x_{1N} \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ x_{p1} & x_{p2} & ... & x_{pN} \end{bmatrix} \tag{6.6}$$

The probability density function of X is given by (Srivastava and Khatri, 1979)

$$\begin{aligned} p(X) &= [(2\pi)^p |V|]^{-\frac{1}{2}N} exp(-\tfrac{1}{2}\textstyle\sum_{i=1}^{N}(x_i - \hat{x})'V^{-1}(x_i - \hat{x})) \\ &= [(2\pi)^p |V|]^{-\frac{1}{2}N} etr[-\tfrac{1}{2}V^{-1}(X - \hat{x}e')(X - \hat{x}e')'] \end{aligned} \tag{6.7}$$

where $(.)'$ denotes the transpose of the matrix or vector, etr denotes the exponential of a trace of a matrix, $e' = (1, 1, ..., 1)$ an N row vector of ones, and the expected value of X, $E(X)$ can be written as a pxn matrix such that

$$E(X) = \begin{bmatrix} \hat{x}_1 & \hat{x}_1 & ... & \hat{x}_1 \\ \hat{x}_2 & \hat{x}_2 & ... & \hat{x}_2 \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ \hat{x}_p & \hat{x}_p & ... & \hat{x}_p \end{bmatrix} = \hat{x}e'. \tag{6.8}$$

Now, for the classification problem, it is known that $\pi_0 = \pi_i$ for i being either 1 or 2. The problem is to find for which i this is true. In the case of multivariate normal populations with common nonsingular covariance matrix V, the problem reduces to that of finding whether $\hat{\pi}_0$ is closest to $\hat{\pi}_1$ or $\hat{\pi}_0$ is closest to $\hat{\pi}_2$, where $\hat{\pi}_0, \hat{\pi}_1$, and $\hat{\pi}_2$ are respectively the mean vectors for the three populations π_0, π_1, and π_2. Thus if an observation x_0 (a subject with measurements x_0 from π_0) is to be classified in π_1 or π_2 when all parameters $\hat{\pi}_1$, $\hat{\pi}_2$, and V are known, we obtain the likelihood ratio as (Srivastava and Khatri, 1979)

$$\begin{aligned} & etr\{-\tfrac{1}{2}V^{-1}[(x_0 - \hat{\pi}_1)(x_0 - \hat{\pi}_1)' - (x_0 - \hat{\pi}_2)(x_0 - \hat{\pi}_2)']\} \\ = \; & etr\{-\tfrac{1}{2}V^{-1}[\hat{\pi}_1\hat{\pi}_1' - 2\hat{\pi}_1 x_0' + 2\hat{\pi}_2 x_0' - \hat{\pi}_2\hat{\pi}_2']\} \\ = \; & exp[(\hat{\pi}_1 - \hat{\pi}_2)'V^{-1}x_0 - \tfrac{1}{2}(\hat{\pi}_1 - \hat{\pi}_2)'V^{-1}(\hat{\pi}_1 + \hat{\pi}_2)]. \end{aligned} \tag{6.9}$$

The best procedure (Srivastava and Khatri, 1979) is to classify x_0 in π_1 or π_2 according to

$$(\hat{\pi}_1 - \hat{\pi}_2)'V^{-1}x_0 - \frac{1}{2}(\hat{\pi}_1 - \hat{\pi}_2)'V^{-1}(\hat{\pi}_1 + \hat{\pi}_2) <> C \tag{6.10}$$

where C is fixed at some specified level. The equation can alternatively be rewritten as follows: classify x_0 in π_1 if

$$(\hat{\pi}_1 - \hat{\pi}_2)'V^{-1}(x_0 - \hat{\pi}_1) > -\frac{1}{2}d_{Mal} + C \tag{6.11}$$

and classify x_0 in π_2 if

$$(\hat{\pi}_1 - \hat{\pi}_2)'V^{-1}(x_0 - \hat{\pi}_2) < \frac{1}{2}d_{Mal} + C \tag{6.12}$$

where

$$d_{Mal} = (\hat{\pi}_1 - \hat{\pi}_2)'V^{-1}(\hat{\pi}_1 - \hat{\pi}_2) \tag{6.13}$$

which is equivalent to the Mahalanobis distance metric described above. Thus the Mahalanobis distance can be used to classify an unknown sample from a normal population into one of many known normal distributed populations.

Problems with the Mahalanobis metric are that, firstly the inverse of a matrix is computationally intensive to calculate and secondly, where the off-diagonal terms are small compared to the diagonal terms (which is the desirable case for feature vectors as this shows that the features are independent of one another) the inverse can have large errors due to small fluctuations brought on by calculation errors such as round-off and estimation inaccuracies. When the covariance matrix cannot be inverted or cannot be easily inverted, the Mahalanobis distance metric weightings can be reduced to the inverse of the diagonal values of the covariance matrix. Thus, each feature component (or feature coefficient) is weighted by its within-word feature variance thereby reducing the Mahalanobis distance to a weighted distance between two vectors. For two vectors, F_1 and F_2, of N dimensions, the reduced Mahalanobis distance is

$$d_{MalRed}(F_1, F_2) = \sum_{i=1}^{N} w(i).(f_1(i) - f_2(i))^2, \tag{6.14}$$

where $w(i)$ is the inverse of the ith diagonal element of the covariance matrix, V_{wi} (Tohkura, 1987).

It seems reasonable to weight the distance with the inverse of the variance. The feature coefficient with the largest variance is spread widely over the feature space and is less likely to contain useful discriminating information. However, one possible problem with the Mahalanobis distance is that the weightings are based on the *within-class* variances only. Thus weightings are based on the variance of the feature within its own class or data; for speech this relates to different versions of the same word and so is also known as the *intra-word* variance. The within-class variance (or intra-word variance) for a word α, represented by V_{wi}^{α}, is calculated as the sum of the squared differences between a reference template, r_α, which represents the mean of the reference templates, and all the reference templates of that word, $b_{j\alpha}$, where $j\alpha$ is the template version of the word α. Thus the variance can be written as,

$$V_{wi}^{\alpha} = \frac{\sum_{j=1}^{\alpha_J}(b_{j\alpha} - r_\alpha)^2}{\alpha_J} \tag{6.15}$$

where the total number of templates for the word α is α_J. This calculation is achieved over the whole word length by using a method of time normalisation such as DTW (discussed in Chapter 5).

Often the coefficient with a large within-class variance also has a large *between-class* variance. The between-class variance for the word α, represented as V_{bw}^{α}, known as the *inter-word* variance for speech recognition, is the distance between a reference template, r_α, of word α and all other (different) word templates $x_{j\beta}$, so that

$$V_{bw}^{\alpha} = \frac{\sum_{j=1}^{\beta_J}(x_{j\beta} - r_\alpha)^2}{\beta_J}. \tag{6.16}$$

where the total number of different word templates is β_J. DTW is also used to calculate the variance across the whole word. If the coefficients have large within-class and between-class variances, both the correct and incorrect word choices are likely to be well spaced from the reference word template. In such cases coefficients should not necessarily be deweighted (with respect to the other coefficients of the vector) by the within-class variance alone as occurs for the Mahalanobis distance. In such cases it may be better to use a weighting based on the ratio *between-class variance/within-class*

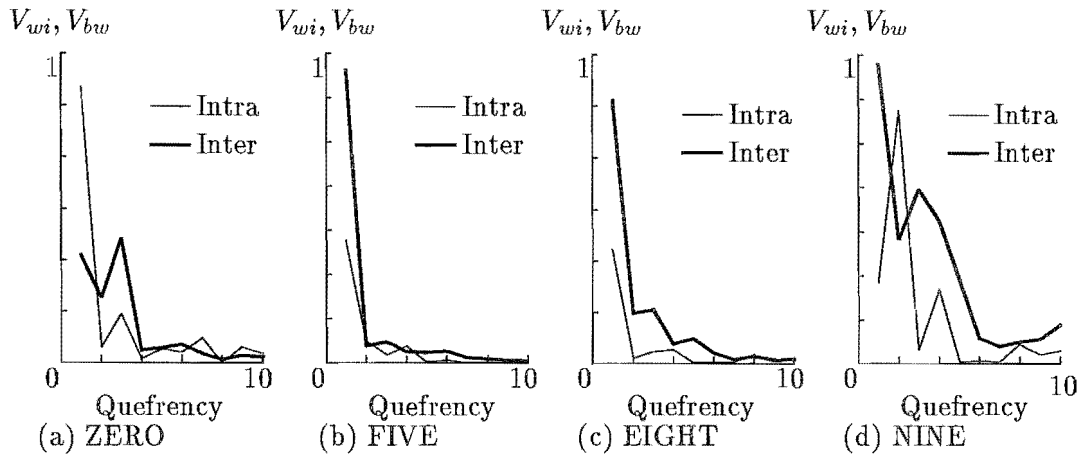


Figure 6.1. Comparison of inter and intra word variances, V_{bw} and V_{wi}, for cepstral coefficients. Plots for the words (a) ZERO, (b) FIVE, (c) EIGHT, (d) NINE.

variance. The inter-word and intra-word variances, V_{wi} and V_{bw} for cepstral coefficients plotted for a selection of words is shown in Fig. 6.1. The intra-word variance is calculated as the variance (equation 6.15) (across a complete word) for each individual coefficient with that same coefficient from other templates of the word from its own word class. The inter-word variance (equation 6.16) is calculated (across a complete word) or each individual coefficient with that same coefficient from another word class. The word were time normalised using a UE21 DTW method with non-constant weightings (refer §5.1.4). An examination of Fig. 6.1 reveals that, for a particular word, often the coefficients with large intra-word variance also have large inter-word variance, for example the first coefficient (quefrency = 1) for the words in 6.1(b) and 6.1(c) have high inter and intra variance and is highlighted by Fig. 6.1(c) which shows features with large within-class, or intra-word, variances are as well spaced from the inter-word distance variance as those features with smaller intra-word variance. Hence coefficients with large intra-word variances are as capable of distinguishing words as those with the small intra-word variances. Although this discussion is only heuristic, testing was not undertaken on the Mahalanobis method to follow up these conclusions. The main reasons why testing was not carried out were, firstly, because the calculation of the Mahalanobis distance is time consuming and this system was constrained to being a real-time one, and secondly, other weighting schemes, such as the RPS, lifters, and quefrency schemes (discussed below) which are faster to calculate have also been shown to give as accurate results as the Mahalanobis (Tohkura, 1987),(Hanson and Wakita, 1986)) and so were considered more useful.

Other weighting schemes which use proportional weightings have been devised that are particularly useful for cepstral coefficients and shown to be equivalent or better than the Mahalanobis method. They are known as the *linear* quefrency weighting and the *root power sum* quefrency weighting (RPS). Weighting is by the coefficient indices or quefrencies (see §4.4) such that low-order coefficients are weighted less than high-order coefficients. The linear quefrency weighted distance between two cepstral vectors, C_1 and C_2, is expressed as

$$d_{QW}(C_1, C_2) = \sum_{i=1}^{p} i.(c_1(i) - c_2(i))^2. \tag{6.17}$$

In contrast with the linear quefrency method the RPS quefrency method uses the square

of the coefficient's index such that the distance between two cepstral vectors, C_1 and C_2 is expressed as

$$d_{RPS}(C_1, C_2) = \sum_{i=1}^{p} (i(c_1(i) - c_2(i)))^2. \tag{6.18}$$

Quefrency weighting schemes are particularly useful for cepstral coefficients because they deweight the lower order coefficients. Deweighting the lower order coefficients reduces the effect of noise on the system ((Gold and Rader, 1969)) as the lower order quefrencies appear to be affected more highly by noise and recording artefacts (Tohkura, 1987). Another reason for weighting the coefficients by their quefrency is due to the reduction of the magnitude of the distances between coefficients at higher orders. Therefore increasing the weighting for higher order cepstral coefficients improves their discriminating ability by normalising the magnitudes of coefficients. These weighting schemes may be more appropriate for recognition particularly under noisy conditions.

6.2.1 Lifters

The weightings of cepstral distance measures fall into a larger category of window weightings for cepstral coefficients known as cepstral *lifters* (Juang *et al.*, 1987; Junqua and Wakita, 1989). Lifters, an anagram of filters, encompass any form of weighting of the cepstral coefficients with a function, $w(i)$. Lifters are applied to the general distance formula between two vectors of cepstral coefficients C_1 and C_2 such that

$$d_{liftered}(C_1, C_2) = \sum_{i=1}^{p} w(i)(c_1(i) - c_2(i))^2. \tag{6.19}$$

Examples of various forms of lifters are presented by Juang *et al*(1987) and shown in Fig. 6.2. Lifters are used to deweight high order cepstral coefficients. Recall that high order coefficients are any coefficients above the order used to calculate the LPC values to derive the cepstral coefficients (refer §4.4). Juang *et al*(1987) claimed that a large part of the variability of the high quefrency terms was 'inherent artefact of the LPC analysis procedure' and so the higher terms are generally not desirable. If weighted distance measures such as RPS and linear quefrency are used the higher order coefficients would be weighted more heavily than the lower order coefficients. The Mahalanobis distance would also weight the higher order cepstral coefficients more heavily since they continue to have reduced within-word variance. Both Tohkura(1986) and Juang *et al*(1987) claim that the higher order coefficients increase recognition error when weighted by the Mahalanobis, RPS quefrency, or linear quefrency distance weightings. However, cepstral weighting lifters where the coefficients are deweighted at both low and high orders is claimed to increase recognition accuracy. Shapes of lifters are based on a trapezoidal, triangular or sine-on-a-pedestal shape (Juang *et al.*, 1987), and a selection of lifters is depicted in Fig. 6.2.

Juang *et al*(1987) noted that the variability of the low quefrency terms can also be attributed to transmission variations, speaker characteristics and vocal effort of the speech. Other effects such as changes of transmission channel roll-off were found to mostly affect the first cepstral coefficient. Another undesirable effect on the low order coefficients is that of *spectral tilt*. Spectral tilt is the effect which occurs with pre-emphasis when high frequency portions of the spectrum are emphasised by increasing the magnitude of this part of the spectrum. Uncontrolled spectral tilt artefacts can be introduced during recording by non-constant recording conditions such as varying microphone and other recording equipment, or by the individual speaker's glottal characteristics. Because these effects can change from recording to recording, they are a disturbing factor in speech recognition.

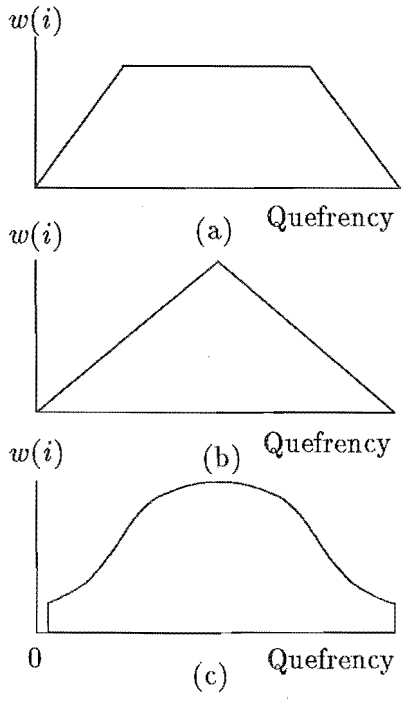


Figure 6.2. A selection of lifters used in weighted cepstral distance measures (a)trapezoid (b)triangular (c) raised sine.

6.2.2 Comparison of Mahalanobis and Quefrency Weighting

Comparisons of the Mahalanobis and RPS quefrency weighted distance measures are discussed in the literature. Tohkura(1987) showed how close the within-class variances of the cepstral coefficients are to being quadratic with respect to the quefrency of the coefficients, showing the similarity between the RPS quefrency and Mahalanobis distance methods. Tohkura's results showed RPS to be only slightly lower in accuracy than the Mahalanobis method. Hanson and Wakita(1986) also compared these two methods and claimed that they are statistically equivalent, since similar accuracies were obtained for both methods.

6.2.3 Comparison of Mahalanobis and Quefrency Weighting with respect to Noise

To test the effect of noise on cepstral coefficients white Gaussian noise was added to the speech and ten cepstral coefficients were computed from ten LPC values. The signal-to-noise ratio (SNR) before the noise was added was 27dB and after the noise, 10dB. The calculation of the SNR is discussed in §8.1.4. The variance of both the within-word and between-word distance for the coefficients was calculated. Fig. 6.3, shows that, generally, both the within-word variance and the between-word variance of the noisy speech drop considerably. If there is a drop in variance, the variance decrease is larger for the lower order coefficients. For such cases, when there is a large amount of noise (SNR=10dB), weighting the coefficients by their inverse variance, as with the Mahalanobis distance measure, would be a mistake since the variance for all coefficients is almost equal. However the low order coefficients are affected more by the noise since they exhibit the greatest change in Fig. 6.3. Using the Mahalanobis distance measure in the noisy case would effectively increase the weighting for the lower order coefficients

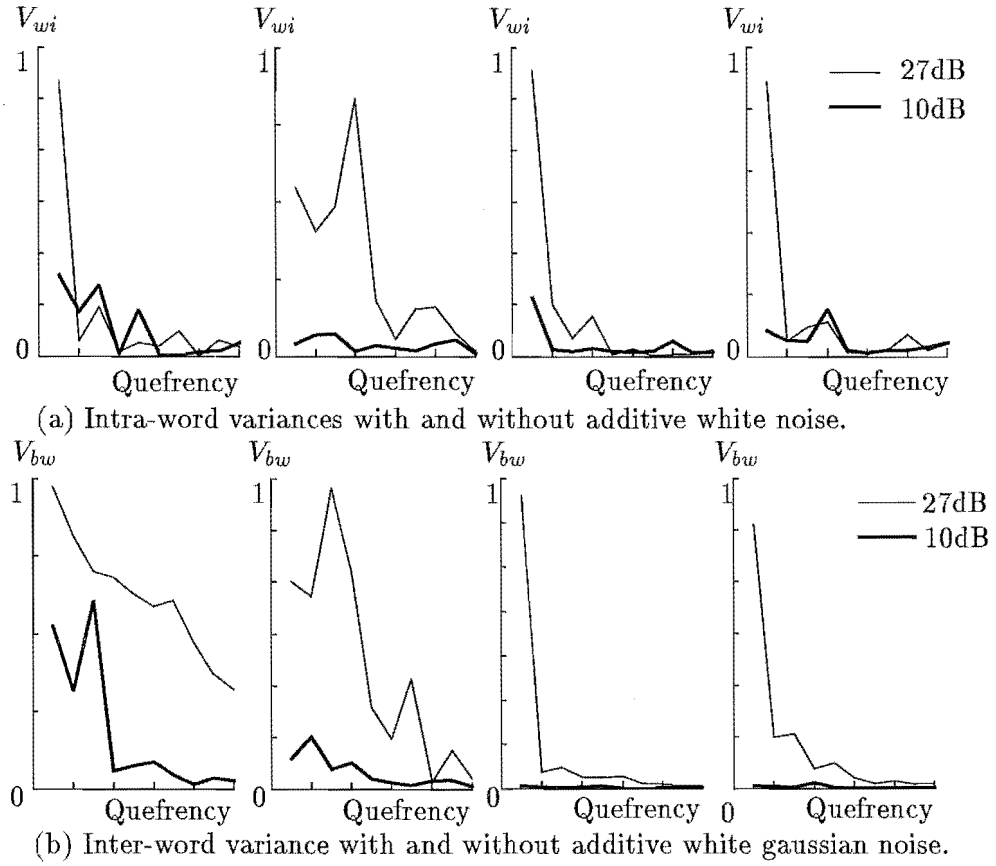


Figure 6.3. The effect of noise on (a)intra-word and (b) inter-word distance variances (d_{s^2}) for the ten cepstral coefficients before and after the addition of random Gaussian noise for the words ZERO, FOUR, SEVEN, NINE.

even though these coefficients are more susceptible to the noise than the higher order coefficients. Thus, for noisy speech, distance measures such as the RPS and linear quefrency would be an advantage as they weight the higher order coefficients more than the lower order coefficients.

Mansour and Juang(1988) examined the effect of noise on the discriminating ability of the Euclidean distance measure for cepstral coefficients. They were particularly interested in the recognition between a clean or noiseless reference and a noisy test. They showed that additive white noise superimposed on the speech reduces the energy (calculated as the norm) of the cepstral vector, if c_0 is removed, and that the reduction of the cepstral norm is directly related to the level of noise added. The reduction in norm is due to the noise flattening the spectrum of the speech and hence reducing the magnitudes of all the cepstral coefficients except the zeroth coefficient which represents the spectral energy. The norm reduction leads to a severe bias in the Euclidean distance calculation and eventually renders this traditional distance measure between noisy and noiseless data useless (Mansour and Juang, 1988). This reduction in cepstral norm has been demonstrated mathematically, (Mansour and Juang, 1988). The cepstral energy (norm), $G(\gamma)$, with additive white noise, where γ represents the additive noise, is calculated from the noisy cepstral coefficients $c_\gamma(i)$ without the zeroth order coefficient

as,

$$G(\gamma) = 2\sum_{i=1}^{\infty} c_\gamma(i)^2. \tag{6.20}$$

The factor of 2 is used because the cepstral function is an even function around the zero point thus, if c_0 is removed, the summation limits can be reduced to begin at 1 not $-\infty$ requiring that the function be doubled.

To show the effect of noise on the norm of the cepstral coefficients, Mansour and Juang related the cepstral energy to the log spectrum of the speech by substituting the all pole model of the frequency spectrum into 6.20, as discussed in §4.4, giving,

$$G(\gamma) = \int_{-\pi}^{\pi} [\ln F_\gamma(w) - \int_{-\pi}^{\pi} \ln F_\gamma(w) \frac{dw}{2\pi}]^2 \frac{dw}{2\pi}. \tag{6.21}$$

where $F_\gamma(w)$ is defined as,

$$F_\gamma(w) = \frac{1}{|A(e^{-jw})|^2} + \gamma \tag{6.22}$$

where $\frac{1}{A(e^{-jw})}$ is the frequency response of an all-pole representation of the vocal tract.

To determine the variation of the norm of the cepstral coefficients with respect to the noise γ, the derivative is taken with respect to γ,

$$dG(\gamma)/d\gamma = 2\int_{-\pi}^{\pi} [lnF_\gamma(w) - \int_{-\pi}^{\pi} lnF_\gamma(w) \frac{dw}{2\pi}] \frac{1}{F_\gamma(w)} \frac{dw}{2\pi}. \tag{6.23}$$

Mansour and Juang show this function has negative values for positive values of γ. This shows the energy in the cepstral domain, without the zeroth quefrency term, is a decreasing function of γ with a maximum at $\gamma = 0$, equivalent to a clean model. A plot of the reduction of the vector norm when the noise is added for a typical word is given in Fig. 6.4, both with (SNR=10dB) and without (SNR=27dB) additive white Gaussian noise.

To attempt to compensate for the mismatch of noise levels that occurs between a clean reference and noisy test templates, Mansour and Juang propose a first order equalisation by multiplying clean reference coefficients by a reduction factor, λ, such that the distance measure between two cepstral vectors C (clean) and C_γ (noisy) is,

$$d(\lambda) = (C_\gamma - \lambda C)^T (C_\gamma - \lambda C), \tag{6.24}$$

where λ is an experimentally derived constant that depends on the noise level and ranges between 1 and 0.27.

Although Mansour and Juang show a problem between noisy and clean speech templates a problem also exists between two noisy speech templates. When comparing the norms of two noisy words the effect of noise may not be considered so important because the norms of both the templates are reduced with the noise. However the magnitude of the norm is not the only attribute affected by noise - the variance of the magnitude is also reduced. This reduction of the variance makes the features of noisy words more alike. It also affects the lower order coefficients more than the higher order coefficients as shown in Fig. 6,3. Thus, for noisy speech, accuracies should be improved by weighting higher order coefficients more heavily. This investigation suggests that, among the Euclidean measures discussed, the linear quefrency or RPS measures should perform best for cepstral coefficients.

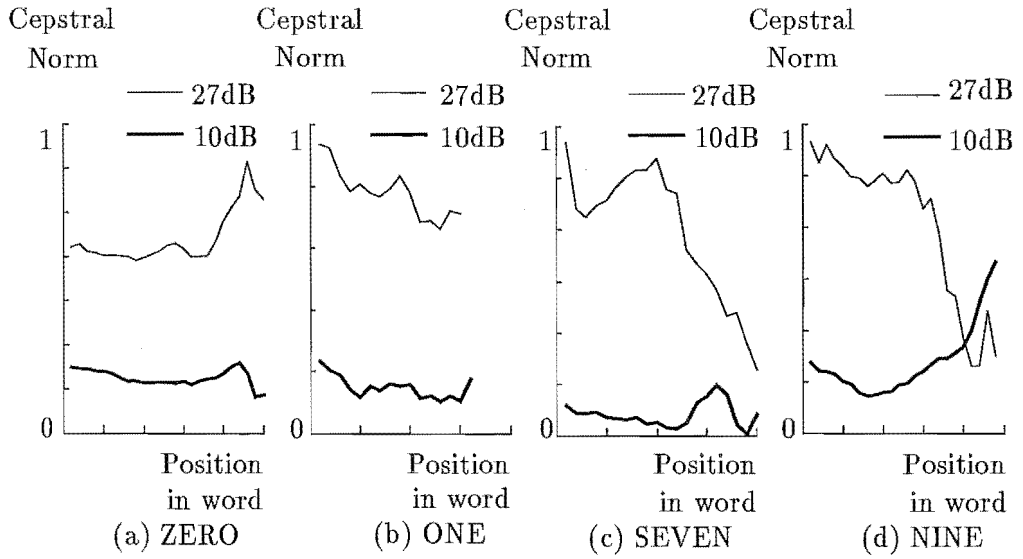


Figure 6.4. Norm of cepstral coefficients across a selection of typical words. Plot shows norm with SNR=27dB and SNR=10dB for the words (a) ZERO, (b) ONE, (c) SEVEN and (d) NINE. Note reduction of both magnitude and variance of norm with greater noise added.

6.3 ANGLE MEASURES

In an attempt to reduce the effect of noise on recognition accuracy Mansour and Juang(1988) began studying the *directional cosine* between cepstral vectors. The directional cosine is the angle between multi-dimensional vectors and for cepstral vectors C_1 and C_2 is defined as,

$$cos\sigma = \frac{C_1^T C_2}{|C_1||C_2|}. \tag{6.25}$$

Mansour *et al*(1988) showed that the directional cosine, or angle, between a clean reference vector, C, and a noisy test vector, C_γ, has a limited deviation that must always be less than or equal to 90 degrees. Thus the variation of the angle between vectors is less sensitive to additive white noise than the norm of the coefficients (refer §6.2) because the variation of the angle is constrained.

Mansour *et al* derived a distance measure using the angle between two vectors. The distance measure was of the form,

$$d_{angle}(C, C_\gamma) = |C_\gamma|^2(1 - cos^2\sigma). \tag{6.26}$$

The family of distance measures which use the angle between two sets of cepstral vectors is known as the cepstral projection measures and can be written as,

$$d_{proj} = |C_\gamma|^\zeta(1 - cos^\zeta\sigma). \tag{6.27}$$

The variation of the distance measure is directly proportional to the angle measurement $(1 - cos^\zeta\sigma)$ and Fig. 6.5 shows the effect of additive white Gaussian noise on the angle multiplier $(1 - cos\sigma)$ in the distance calculation of equation (6.19), when $\zeta = 1$. It can be seen that the magnitude and variation of the angle are not affected as greatly by the noise as was the norm and so making the angle distance measure a better choice than the norm.

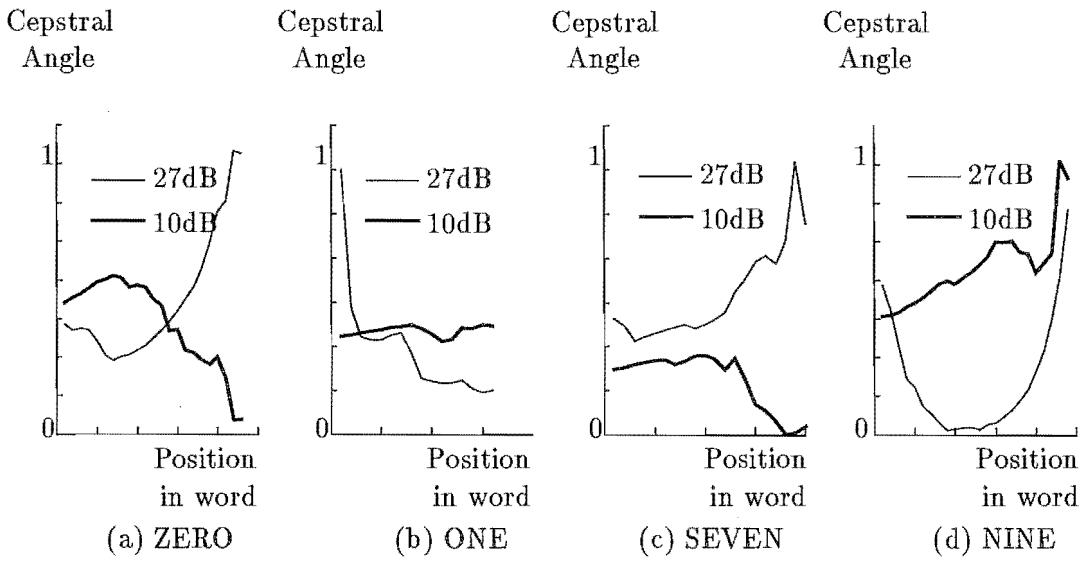


Figure 6.5. Angle of cepstral coefficients across a selection of typical words. Plot shows angle measurement with SNR=27dB and SNR=10dB for words (a) ZERO, (b) ONE, (c) SEVEN and (d) NINE.

6.4 PROBABILITY DISTORTION MEASURES

In this section the likelihood and the log likelihood distance measures are discussed. These measures have the unique property that they are asymmetric, non-linear, and are perceptually meaningful. It has been shown that humans perceive frequency movements of the formant frequencies in a non-linear manner (Bladon, 1985) and that this perceptual effect is not well represented in a Euclidean space. A group of distortion measures which are considered perceptually relevant are the Itakura-Saito maximum likelihood measure, the likelihood measures and the log likelihood measure.

The maximum likelihood measure was initially formulated by Itakura and Saito(1968) who successfully used it in a linear predictive coding (LPC) algorithm for speech. The assumptions required to obtain LPC coefficients from a maximum likelihood formulation are discussed in §4.3. The main assumption is the requirement that speech is a Gaussian process, generated by passing white noise through an all-pole filter of order p. Itakura and Saito(1968) showed that the maximisation of the likelihood formulation for LPC coding was equivalent to minimising their (Itakura-Saito) distortion measure. This measure forms the basis of the spectral matching properties of linear prediction showing why poles are weighted more heavily than zeros (Markel and A.H. Gray, 1976). Mathematically, the maximum likelihood measure can be written as the distance, $d_{IS}(S_1(w), S_2(w))$ between two all-pole power spectra, $S_1(w)$, and $S_2(w)$, where $S_1(w)$ is the short-time spectral density of an test speech signal and $S_2(w)$ is the short-time spectral density of a reference speech signal,

$$d_{IS}(S_1(w), S_2(w)) = \int_{-\pi}^{\pi} \left[\frac{S_1(w)}{S_2(w)} - ln \frac{S_1(w)}{S_2(w)} - 1 \right] \frac{dw}{2\pi}. \tag{6.28}$$

These speech signals are assumed to have been produced by an all-pole model and hence can be represented as,

$$S(w) = \frac{\sigma^2}{|A(e^{jw})|^2}, \tag{6.29}$$

which is the spectral density function of a corresponding pth order all-pole model, with σ representing the gain of the model and $A(e^{jw})$ representing the all-pole filter (Gray

et al., 1980).

The Itakura-Saito measure, d_{IS}, has some interesting characteristics. First it is asymmetric. To understand the nature of the asymmetry of the measure it is usually easier to write it as,

$$d_{IS}(S_1, S_2) = \int_{-\pi}^{\pi} \left[e^{d_{ln}} - d_{ln} - 1 \right] \frac{dw}{2\pi}, \tag{6.30}$$

where,

$$d_{ln} = ln \frac{S_1(w)}{S_2(w)}. \tag{6.31}$$

d_{ln} is itself a distortion measure for speech, giving the difference of log spectra.

The asymmetry with respect to d_{ln}, violates the symmetric requirements of a distortion measure. Examining $e^{d_{ln}} - d_{ln} - 1$ shows the positive values of d_{ln} are more important in the distance measure than the negative values. That is, when the input or test speech $S_1(w)$ is greater than the reference speech $S_2(w)$, the distance grows large (reaching exponential growth), and, with an LPC analysis, the spectra are forced to match closely. In fact the larger d_{ln}, (that is the greater $S_1(w)$ is over $S_2(w)$) the closer $e^{d_{ln}} - d_{ln} - 1$ comes to exponentially increasing. When the input or test speech is smaller than the reference speech the movement of $e^{d_{ln}} - d_{ln} - 1$ with respect to d_{ln} becomes linear and so there is only a small affect on the distance measure (smaller than when the test speech is larger than the reference speech).

Although useful as a perceptually meaningful measure, the Itakura-Saito distortion measure contains the gain of the all-pole model which makes it sensitive to changes in the energy and inappropriate for recognition. This gain term represents the energy of the filter and varies in proportion to the changes in energy in the speaker's voice. The gain term in the d_{IS} measure can be separated from the filter, $S(w)$, by expanding d_{IS} as,

$$d_{IS}(S_1(w), S_2(w)) = (\sigma_1)/(\sigma_2) \int_{-\pi}^{\pi} \exp[d_{ln}] \frac{dw}{2\pi} - ln(\sigma_1)/(\sigma_2) - 1, \tag{6.32}$$

where σ_1 is the gain of the input signal and σ_2 is the gain of the reference signal and

$$\int_{-\pi}^{\pi} ln \frac{\sigma^2}{|A(e^{jw})|^2} \frac{dw}{2\pi} = ln(\sigma^2). \tag{6.33}$$

To remove the effect of gain on this measure Itakura later proposed a *gain-normalised* measure, also known as the *likelihood ratio*. This measure is calculated from the Itakura-Saito measure by setting the all-pole model gain of two systems to be equal, which is equivalent to normalising the spectra of the two systems. This reduces the distortion measure of equation (6.24) to,

$$D_{LR}(S_1(w), S_2(w)) = \int_{-\pi}^{\pi} exp[d_{ln}] \frac{dw}{2\pi} - 1, \tag{6.34}$$

and by substituting equations (6.23) and (6.21)

$$\begin{aligned} &= \int_{-\pi}^{\pi} \left[\frac{S_1(w)}{S_2(w)} \right] \frac{dw}{2\pi} - 1, \\ &= \int_{-\pi}^{\pi} \left| \frac{A_2(e^{jw})}{A_1(e^{jw})} \right|^2 \frac{dw}{2\pi} - 1, \end{aligned} \tag{6.35}$$

where $A_1(e^{jw})$ and $A_2(e^{jw})$ are the all-pole representations of the frequency spectrums or LPC models as introduced in equation (6.21). Equation 6.35 is a useful

distortion measure if the recognition features used are LPC coefficients. In such cases $D_{LR}(S_1(w), S_2(w))$ can be further reduced and calculated as,

$$D_{LR}(S_1(w), S_2(w)) = \frac{a_2^T \mathbf{R}_1 a_2}{a_1^T \mathbf{R}_1 a_1} - 1, \tag{6.36}$$

where $\mathbf{R}_1$ is a matrix containing the first $p+1$ autocorrelation terms of the test word, and a_1, a_2 are LPC coefficient vectors (pth order) of the test and reference words respectively. In this form the distance measure can be computed in a straight forward manner since $a_1^T \mathbf{R}_1 a_1$ is the prediction error, α_1, (refer §4.3) calculated at the same time as the LPC coefficients. Thus the distance measure can be written as,

$$D_{LR}(S_1(w), S_2(w)) = \frac{a_2^T \mathbf{R}_1 a_2}{\alpha_1} - 1. \tag{6.37}$$

Itakura proposed a log version of the distortion measure known as the *log likelihood ratio* and is

$$D_{LLR}(S_1(w), S_2(w)) = \ln \left[\frac{a_2^T \mathbf{R}_1 a_2}{\alpha_1} \right]. \tag{6.38}$$

The above equations have a term $a_2^T \mathbf{R}_1 a_2$ which can be considered an error term. Letting $\alpha_2 = a_2^T R_1 a_2$, the log likelihood distance measure can be written as,

$$D_{LLR}(S_1(w), S_2(w)) = \ln \left[\frac{\alpha_2}{\alpha_1} \right]. \tag{6.39}$$

This distance measure can be regarded as a ratio of two error signals. The first error α_1 being the residual energy when a frame of test signal is passed through an LPC filter optimised for a frame of test data. The second error, α_2, is the optimal residual energy found when a frame of the reference signal is passed through an LPC filter optimised for a frame of test data.

6.5 DISTANCE MEASURES USED IN EXPERIMENTS

The distance measures chosen for feature testing in Chapter 8 are selected from those discussed in the previous sections. A constraint on the distance measures chosen was that they are required to be calculated within the limitations of real-time operation. The following section discusses those distance measures chosen and the reasons why.

The first measure tested was the unity weighted Euclidean distance which was tested on all features because of its ease of calculation and because it requires few calculations. Further, the Euclidean distance is frequently used in word recognition tests and can be treated as a reference for discussing the recognition accuracies of other methods. For cepstral coefficients the unity weighted Euclidean distance measure is not considered the optimum distance measure so cepstral coefficients were also tested with both the quefrency weighted Euclidean distance measure and the projection distance measure. The Mahalanobis distance measure was not chosen because of the difficulty in calculating the inverse covariance matrix and because it was considered to be only as good as the weighted cepstral measure shown in the discussion in §6.2.2. Probability distortion measures were not chosen for two reasons. First, they require too many calculations to operate within the real-time constraint and secondly, initial testing with the log likelihood method gave very low accuracies (<50%) and so was not considered further.

The formula for each of the distance measures selected is given in Table 6.1.

Distance Measure	Title	Formula
Unity weighted Euclidean	d_{Euclid}	$\sum_{i=1}^{N}(f_1(i)-f_2(i))^2$
Quefrency weighted Euclidean	d_{we}	$\sum_{i=1}^{N} w(i)(f_1(i)-f_2(i))^2$
Projection distance	d_{proj}	$\|C_2\|^2(1-cos\sigma)$

Table 6.1. Distance measures used in word recognition experiments.

Chapter 7

AN EXAMINATION OF THE PROCEDURES FOR THE RECOGNITION ALGORITHM

This chapter details the algorithms implemented to perform isolated-word recognition. These algorithms are used in speaker-dependent and speaker-independent feature comparison tests described in Chapter 8.

A block diagram of the complete recognition algorithm is shown in Fig. 7.1. The individual algorithms, illustrated as separate blocks of Fig. 7.1, include methods for performing such operations as word endpoint detection, feature extraction, training the system (such as clustering), and time alignment using DTW, and are discussed generally in Chapters 4, 5, and 6. This chapter discusses the implementation details of the specific methods used in the experimental work reported in Chapter 8.

To test the various parameters associated with each block of Fig. 7.1 preliminary recognition experiments were undertaken. These recognition experiments are described fully in this chapter.

7.1 ENDPOINT DETECTION

This section discusses the methods used for the automatic detection of endpoints (*endpointing*) of reference and test words. Methods of automatic endpoint detection involve using some feature, or group of features, to separate spoken words from background noise. This is usually achieved by detecting the difference in energy or spectra be-

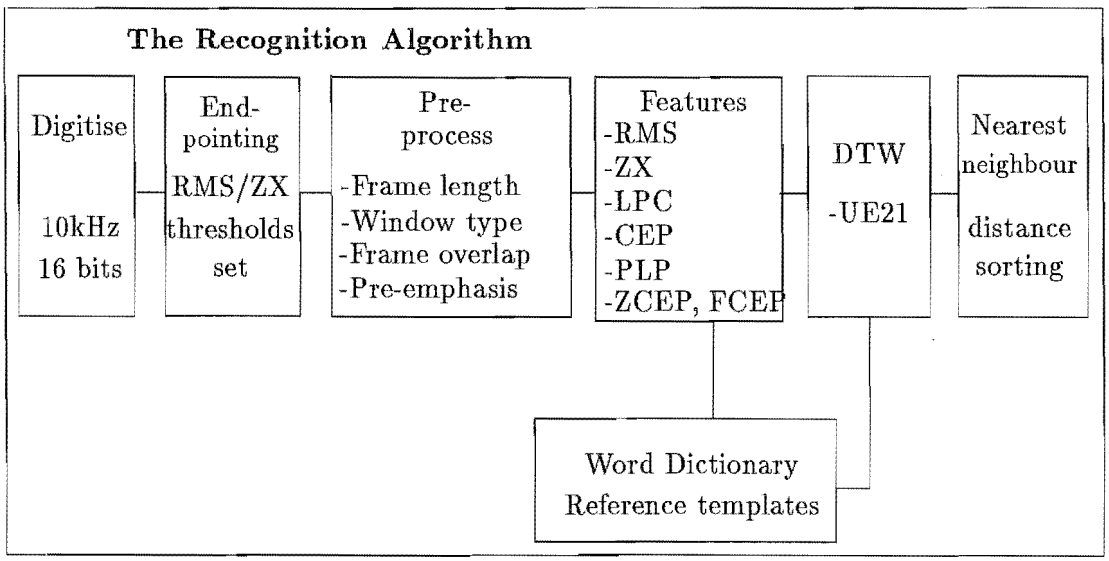


Figure 7.1. Block diagram of recognition scheme designed for testing various features and variables

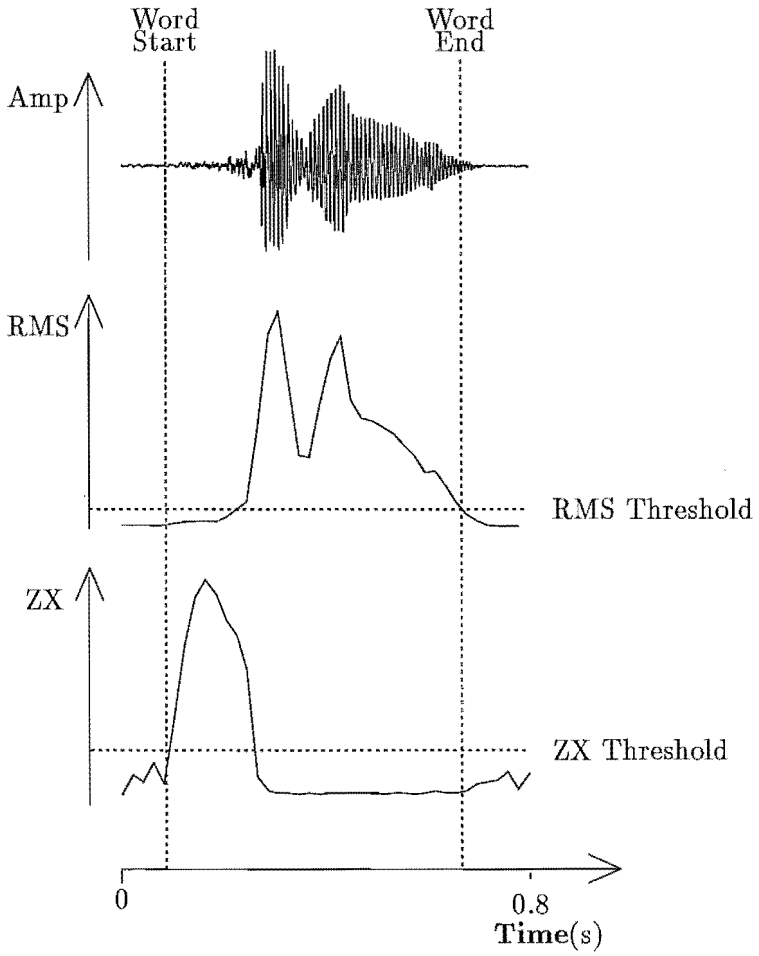


Figure 7.2. Example of endpointing using energy and zero-crossing. Energy and zero-crossing thresholds are set prior to detection.

tween background noise and speech. Rabiner(1978) discusses one of the most reliable methods which uses two simple time-domain measurements; energy (refer §4.1) and zero-crossing rate (ZX) (refer §4.2). In this thesis a method similar to that discussed by Rabiner(1978) is used. RMS intensity (refer §4.1) and ZX values of words are extracted and endpointing rejection thresholds are set to discriminate between silence and speech. The method requires that both the ZX and RMS values are monitored until either goes above their individual set thresholds. Both ZX and RMS are needed as ZX is the feature used to determine unvioced sounds from silence, as both these can have similar RMS values (refer Fig. 7.2 at the beginning of the word), while RMS is used to determine voiced sounds from noise as voiced sounds have a much higher RMS level.

The word is deemed to have begun if either the ZX or RMS values continue to stay above the threshold level for a required number of frames (usually 4-5 frames). The word starting position is taken at the position at which the first frame of data went above the threshold. The word ending position is calculated in a similar way except that both the ZX and RMS values are monitored for the time at which both drop below their threshold value. This point is taken as the endpoint of the word only if the ZX and RMS values stay below the threshold for a set period of time (usually 4-5 frames).

An illustration of how speech endpoints are determined from the thresholds is shown

in Fig. 7.2. These thresholds are set prior to endpoint and word discrimination and are not changed with respect to speaker vocal intensity or background noise. Since the rejection thresholds are preset, that is they are set before speaking levels are known, the rejection thresholds have to be set low enough to ensure that every word is recorded in its entirety and generally this allows one or two frames of non-speech noise, such as breath noise or background noise, to occur at the beginning and ending of each word. However, it was considered better to allow noise and artefacts in the speech by using low (or relaxed) RMS thresholds than to possibly cut off beginnings and endings of the word by using high RMS thresholds. If parts of the speech were cut off important information could be lost, which could, for example, make a word such as NINE sound more like the word ONE. Relaxing endpoint constraints by lowering endpointing thresholds and causing endpointing errors, may not cause too large a problem for the recognition system because it was expected that noise at the beginning and end of the words would be largely negated during the recognition phase for two reasons. Firstly, because most of the words are endpointed to incorporate some noise at the beginning and end of words, during the matching phase noise of a test word should be matched to noise contained in the reference words. Secondly, although the number of frames of noise at the ends of the words is unknown a flexible alignment method such as the UE21 DTW (refer §5.1.2.5) method was used so that the beginnings and endings of words are not severely constrained during recognition. This method more effectively allows the alignment of noise to noise and speech to speech attempting to correct for framing errors (and for this case 3 frames was chosen). An example of the errors in endpoint detection (causing framing errors) corrected during the recognition operation is illustrated in Fig. 7.3. A perfect alignment of different words cannot, however, be expected because often the features are not significantly different between noise and speech and between different frames of noise and noise, or speech and speech.

7.2 OPTIMIZATION OF RMS THRESHOLD

Although both ZX and RMS were used to distinguish sounds from silence, the main reason for using zero crossing measurements was to distinguish unvoiced sounds from background noise. Hence a ZX threshold level could be simply found by examining plots of the speech and also by examining the pdf plots drawn in Fig. 4.4 and this has been discussed in 7.3. As can be seen from Fig. 7.2 the position where background noise ends and unvoiced sounds begin produces a definite and sudden change in the ZXR, hence an accurate position can be located. An optimum RMS level, however, was more difficult to determine. Examining Fig. 7.2, with the RMS threshold set as shown, the actual ending of the word looks to be slightly past the calculated endpoint. Also the determination of the point where background noise ends and sounds begin (or vice versa) is quite subjective, it was not known whether it would be better (for recognition accuracy) for an endpointer to be highly accurate or not. It was decided to run tests on various RMS levels to determine an optimum RMS level, that is one which produced high recognition accuracy even if this did not produce a high endpoint accuracy.

The method of examining whether background noise at the beginning and ending of the words would increase or decrease recognition accuracy was to test different thresholds with recognition trials. To allow for testing the RMS levels the tests were undertaken by endpointing using only the RMS threshold. Thus a direct link between threshold level and speech amplitude could be found. A large RMS value would remove all silence and possibly some of the word. A low RMS threshold value would allow noise to exist at the beginnings and endings of words. RMS was the only variable tested (rather than both RMS and ZX) so that it would be clear that the effect being

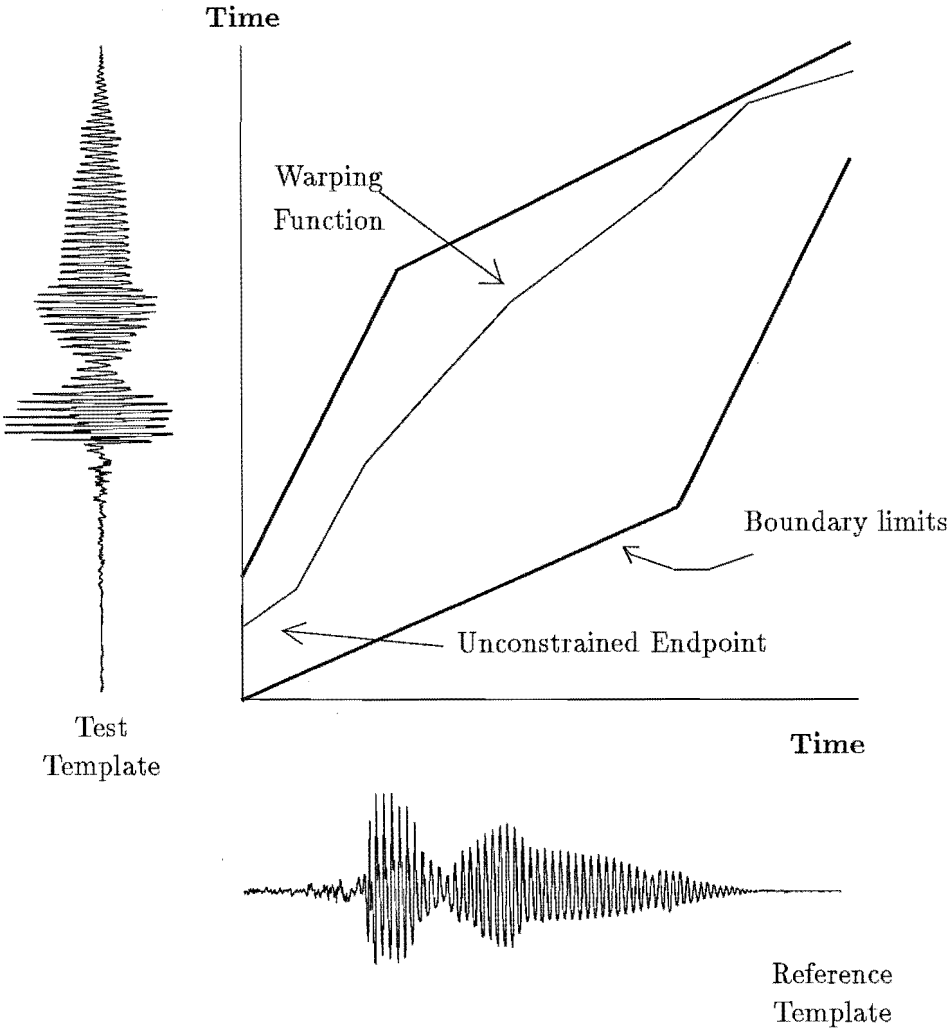


Figure 7.3. Effect of endpoint mismatch occurring due to endpointing error. Attempted endpointing alignment correction can occur with a UE21 DTW operation due to variable alignment of unconstrained endpoint method. Alignment error, as shown at the beginning of the warping path may still occur due to local path weightings (refer §5.1.2.3)

tested was that only of signal levels and not signal frequencies. RMS thresholds of 100, 200, and 300 was used because the RMS levels for background noise ranged between 20 and 200 for frames of 200 samples. Accuracy results for the male and female speakers are given in Fig. 7.4 using the various endpointing RMS thresholds of 100, 200 and 300 and calculated for windows of 200 samples (sampled at 10 kHz) with a noise level as discussed in §8.1.4. For these trials both male and female speakers were tested in speaker-dependent mode for a range of features (as discussed in Chapter 4). Ten reference templates per word were used. For the recognition tests a UE21 method was used with constant weightings (refer §5.1). The distance measure was the Euclidean distance measure and this was used for all features during the recognition tests. The accuracies plotted in Fig. 7.4 are averaged over all the features (RMS, ZX, LPC, CEP, ZCEP, FCEP and PLP) and represents the general trend for each feature.

Results from these tests show that as the RMS threshold increases the recognition accuracies decrease. For a threshold of 100 the average accuracies of the male and female speakers was 51%. Increasing the RMS threshold to 300, decreases the average

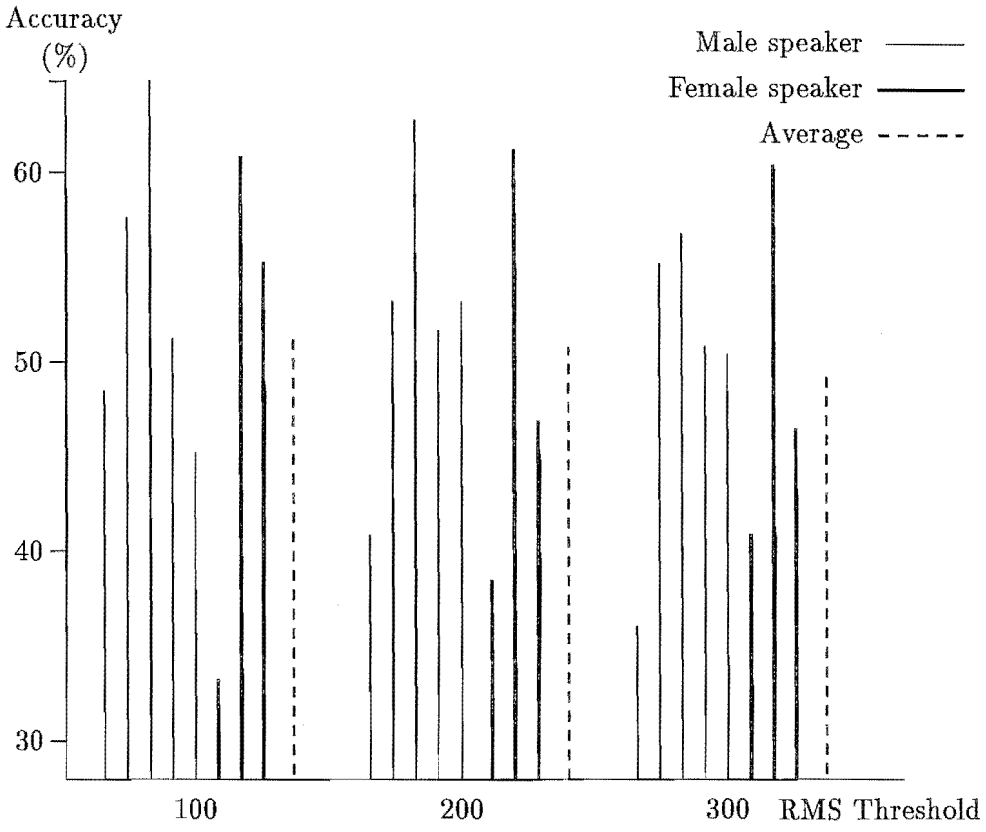


Figure 7.4. The effect of increasing RMS threshold (100,200, and 300) for endpointing on the recognition accuracy. Recognition trials for five male and three female New Zealand speakers averaged over range of features including RMS, ZX, LPC, CEP, ZCEP, FCEP, and PLP (refer Chapter 4). A UE21 DTW method with constant weightings was used. Vocabulary consisted of the word digits ZERO through NINE. Ten templates were stored for each word.

accuracies for the male and female speakers to 49%. Examination of the individual speakers accuracies showed that 6 out of 8 decreased, following the average trend. However, 2 of the 8 speakers tested had accuracy increases. The accuracy increases were not gender specific since they occurred for one male and one female speaker. It is more likely that the different accuracies are related to the individual speaker's speech energy and the noise level. Although the recognition accuracies between the three threshold levels are different, the differences are not significant (based on a paired-t test at the 95% level). Because there is a slight trend to higher accuracies with lower RMS threshold values, however, it was decided to keep the endpointing RMS threshold values at 100 (for frame sizes of 200 samples). The testing discussed in the above paragraphs was solely to determine whether noise at the beginning and ending of the words would degrade the accuracy. However, endpoint detection for all other recognition tests discussed in this thesis used both RMS and ZX to allow the discrimination of unvoiced sounds.

7.3 OPTIMIZATION OF ZX THRESHOLD

As discussed above the determination of the optimum ZX level was achieved by sighting the words and examining the unvoiced sounds ZX pdf. The ZX level chosen was

normalised to the length of the data frame size. The ZX level chosen was generally above the background noise ZX value and therefore the addition of noise at the beginning and endings of the speech templates was solely dependent on endpointing errors due to the RMS level. There were, however, some noise spikes during the silence (refer §8.1.3) which would have produced ZX rates momentarily higher than the threshold used. These aberrations, however, were generally removed when the sentences were edited to words. A value of 35 for the ZX threshold value was chosen for a 200 sample frame size.

7.4 METHODS OF PRE-PROCESSING AND FEATURE EXTRACTION

This section outlines the methods used for the extraction of features tested in Chapter 8. The algorithm used to facilitate feature selection for the experimental recognition scheme, discussed in Chapter 8, was made as general as possible so that any feature to be tested could be chosen with a range of pre-processing techniques.

Pre-processing techniques are defined in this thesis as any technique that can be varied for all features and chosen prior to the feature extraction. The pre-processing techniques include choice of frame size, length of frame overlap, pre-emphasis and type of windowing.

To perform feature extraction the speech is first digitised (refer §8.1.3) and endpointed (refer §7.1). The speech can then be blocked into lengths of various frame sizes. Pre-processing variables such as frame overlap, pre-emphasis and the type of windowing can then be chosen. Features are then extracted and saved in feature files to be used in recognition experiments.

7.4.1 The Features Extracted

The features tested and discussed generally in Chapter 4, were root-mean-squared intensity (RMS), zero-crossings (ZX), linear prediction coding (LPC), perceptual linear predictors (PLP), cepstral coefficients (CEP), and transitional cepstral coefficients (ZCEP, FCEP). The following specifies the methods used for the extraction of these features for the recognition experiments discussed in Chapter 8.

RMS (intensity) was calculated as discussed in §4.1.

Zero-crossings were calculated as a count of the number of times the waveform crossed the zero axis in a frame length. No thresholding or offsets were added to remove effects of noise on the zero-crossing count.

LPC coefficients were calculated using the autocorrelation method. The Durbin-Levinson method of recursion was used to calculate ten coefficients from eleven autocorrelation coefficients (Rabiner and Schafer, 1978).

Ten cepstral coefficients were calculated from the ten LPCs using the recursive formula discussed in §4.4.

Transitional data was calculated from the cepstral coefficients. Both the zeroth and first order transitional coefficients were calculated as outlined in §4.5. The number of frames over which the transitional data is calculated (the *frame-width*) and the transitional cepstral window weighting used during calculations were investigated (these were discussed in §4.5. Testing using a UE21 DTW method, with constant weightings (refer 5.1.2.4) and employing one male and one female speaker on the digit vocabulary, was undertaken to select the optimum number required for the frame-width (where a frame of data is 200 samples) and the optimum window weighting. Results of these tests are given in Table 7.1. Examining the results of Table 7.1 showed that average

accuracies improved as the frame-width increased up to a maximum of 5 frames wide and dropped beyond this frame-width. At a frame-width of 5 the accuracies for first and zeroth order transitional cepstral (linear window) were 88% and 96% respectively. Because highest average accuracies were obtained for a frame-width of five frames along with a linear weighting, the transitional data features were calculated with these variables set as such.

Number of Frames	Window weighting ((frame-width)	First order transitional coeff accuracy(%)	Zeroth order transitional coeff accuracy(%)	Mean accuracy
3	none	87	90	88.5
	linear	79	91	85.0
	square	77	-	-
5	none	86	94	90.0
	linear	88	96	92.0
	square	80	-	-
7	none	85	94	89.5
	linear	88	95	91.5
	square	91	91	91
9	none	87	92	89.5
	linear	88	95	91.5
	square	91	90	90.5

Table 7.1. An examination of various parameters used when calculating transitional cepstral features. Recognition results are given for different frame-widths and different windows. Recognition is for one male and one female speaker, using the digit vocabulary. Ten reference templates of each word are stored and a UE21 DTW method of comparison is used with ten test words for each digit for each speaker.

The calculation of PLPs follows exactly that of Hermansky(1990) as outlined in §4.6.

7.5 TRAINING OF REFERENCE TEMPLATES

The training algorithm used in the recogniser is important because it affects the selection of reference templates stored. It is important to have a 'good' selection of reference templates which represent the variation in the speaker's characteristics across the words spoken. Correctly representing the spread of a speaker's utterances is important to accurately characterise the variations of that speaker's voicing of the utterances. In this section two methods of training the recogniser with reference templates are discussed. These methods are evaluated in series of recognition tests and results are discussed in Chapter 8.

The first training method represents the speaker's voice by storing many representations of a speaker voicing the required word. This method assumes that the speaker's voice can be characterised by a predetermined number of templates and each spoken template is stored for reference. Usually the number of representations stored is be-

tween 1 and 10. This method, known as the *casual training method* (Rabiner and Schmidt, 1980), is simple to implement and therefore commonly used (Itakura, 1974; Rosenberg and Itakura, 1976).

The second method, known as the *statistical clustering method* (Rabiner and Schmidt, 1980), is somewhat more complicated. As with the casual training method this method also assumes that a speaker's voice can be properly represented by obtaining enough representations of a speaker voicing the required word. Usually between 10 and 100 representations are recorded. This method differs from the casual training method in that a reduction in the number of templates is achieved by clustering these representations. A 'good' clustering procedure for a word recognition scheme performs a reduction in the number of templates modelling a word without affecting the recognition accuracy for a word. Template reduction, via clustering, is usually achieved by obtaining an average of the features of the time-aligned patterns within each cluster. The average template then represents the clustered templates.

In the experiments reported here, clustering is performed by averaging the data which are most similar. Data similarity is calculated using a distance score, discussed in Chapter 6. The clustering method in this thesis follows an agglomerative method (Everitt, 1980) as explained in the steps below but with distances (or similarity) scores calculated using a DTW. Word averages for clustered data are produced by performing averaging of feature sets along the calculated warping path. The statistical clustering method algorithm used in this thesis can be expanded as;

1 **Set** parameters for the clustering procedure. The parameters are the required number of clusters, the type of features and the distance measure;

2 **Read** all reference templates of a particular word to be clustered, $word_1, word_2..word_N$, where N is the number of original template representations and is usually between 10 to 100 templates;

3 **Calculate** distances between every pair of words using the UE21 DTW algorithm. Distances must be calculated between all word combinations because the DTW algorithm is asymmetric (refer 5.1.2.2). The asymmetry means that the distance D_{12} between $word_1$ and $word_2$ is not the same as the distance D_{21}, the distance between $word_2$ and $word_1$. The distances for the same words pairs, D_{12} and D_{21} are averaged so that an average distance is found for each word pair. When length variations between templates exceed global constraint variations (refer §5.1.2.5) global distances between words cannot be computed. Thus, it is sometimes not possible to form the number of clustered templates specified, if this is the case the algorithm is stopped (see step 7);

4 **Find** the word pair with the smallest distance;

5 **Average** the feature sets of the two templates with the minimum word pair distance. Features are aligned for averaging by tracing back along the warping path calculated during DTW calculations in Step 3. Remove the two words, now averaged, from the word set, replace by averaged word representation;

6 **Reset** algorithm parameters setting total number of templates $N \leftarrow N - 1$;

7 **Repeat** Steps 3 to 6 until number of templates N is equal to number of clusters required, or until a minimum word pair distances cannot be found due to global constraints.

Both the casual training method and the statistical clustering method are investigated in Chapter 8. For the casual training method accuracies with 10, 6 and 2

templates are investigated. The templates were chosen at random from the 20 original templates however the reference and test templates were changed for each new test, this method, known as the Jackknife method, is further discussed in §8.1.5. For the statistical clustering method 6 and 2 templates are formed, where possible, from 10 original templates.

7.6 THE DYNAMIC TIME WARPING (DTW) ALGORITHM

The following sections discuss those variables in the DTW algorithm which are, of necessity, set prior to recognition. These include the local constraints, the global constraints and the search techniques. These variables are also discussed, more generally, in §5.1.2.2, §5.1.4, §5.1.2.4 and §5.1.6.

7.6.1 Global constraints

To test whether the global constraints of the DTW algorithm, as discussed in §5.1.2.2, have any significant effect on the accuracy, and so to choose the best constraints to use, three types of global constraints were tested; the constrained endpoint 2-to-1 method (CE21), the unconstrained endpoint 2-to-1 method (UE21) and the unconstrained endpoint band method (UEB). These three methods were previewed in §5.1.4.

A preliminary recognition test based on the speech of one NZ male and one NZ female speaker (in speaker dependent mode) was carried out to gain insight into the relative accuracies of the three DTW methods. The recognition tests undertaken used the extracted features of linear prediction coefficients (LPC), cepstral coefficients (CEP) and zeroth order transitional cepstrals (ZCEP), as discussed in §4.3, §4.4 and §4.5 respectively. A vocabulary consisting of 12 repetitions of the words ZERO through NINE was used with two utterances of each word yielding reference templates. A Euclidean distance was chosen to measure the distance between reference and test templates. The results of testing the three DTW methods in this way are depicted in Fig. 7.5. The results shown for the UEB method in Fig. 7.5 used a 13 frame width band, but as shown in Fig. 7.6, accuracy for this method varies considerably with band width. A 13 frame width band generally gives the highest accuracies; band widths above this level were not tested because the DTW method became too slow (refer Fig. 7.7). The results of Fig. 7.5 show that the UE21 method gives significantly higher accuracies than the CE21 method (significant at the 95% level using a paired-t test). It is also noted from Fig. 7.5 that the UEB method is a significant improvement over the UE21 method (at the 95% level using a paired-t test). From an examination of the individual features it is found that this accuracy improvement seems to be predominantly for LPCs rather than the other features tested.

Although the UEB method gave higher recognition rates than both the CE21 and the UE21 methods it was not used for the isolated word recognition tests reported in Chapter 8. The reason for this is that the band DTW method was not considered until after the UE21 and CE21 methods had already been operating and extensively tested. However from these results it appears that the band DTW method would give higher word recognition accuracies. The band DTW method was considered only for the continuous recognition scheme discussed in Chapter 9.

7.6.2 Local Constraints

Weightings on the local path movements (local path constraints), as discussed in §5.1.2.4, were tested to determine whether they had a significant effect on recogni-

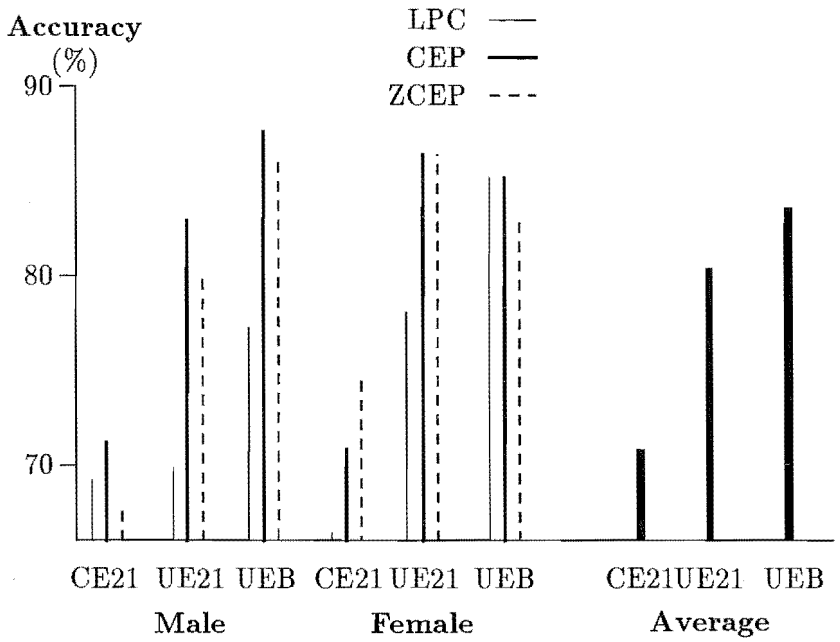


Figure 7.5. Accuracies for recognition test with three different types of DTW schemes. Recognition tests on both male and female speakers, speaker dependent for the 10 word vocabulary ZERO to NINE. The tests used two reference utterances per word selected using the casual training method. The recognition was undertaken using a Euclidean distance for the three features; linear prediction coefficients (LPC), cepstral coefficients (CEP), and zeroth order transitional cepstral coefficients (ZCEP).

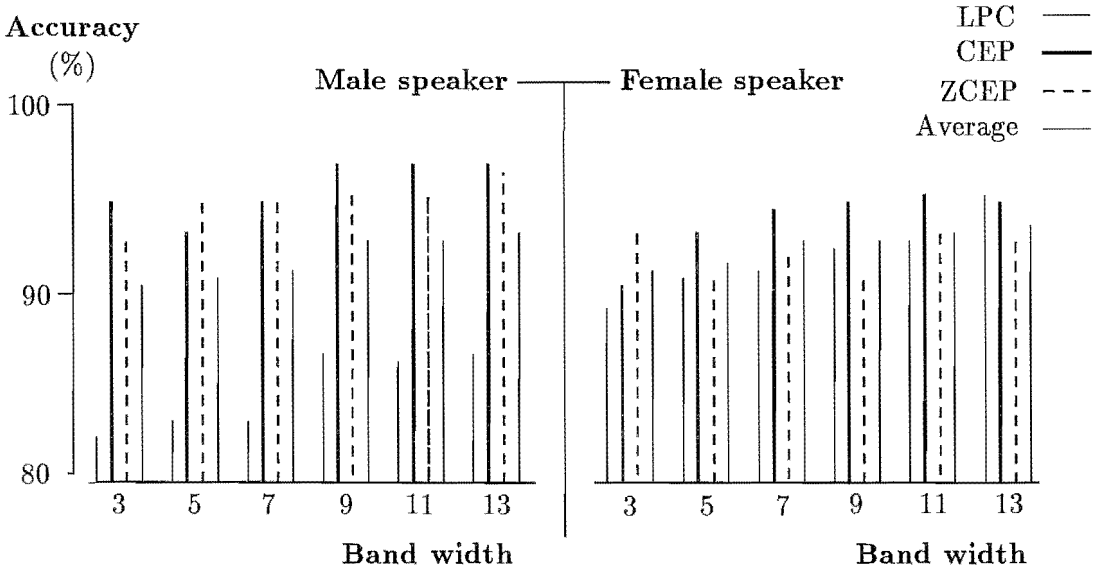


Figure 7.6. Accuracy of the UEB DTW method with respect to the width of the band constraint, which was set to be either 3,5,7,9,11 or 13 frames wide. Recognition tests were undertaken with one male and one female speaker on the ten word vocabulary ZERO to NINE. A Euclidean distance measure was used. The features tested were linear prediction coefficients (LPC), cepstral coefficients (CEP) and zeroth order transitional cepstral coefficients (ZCEP). Two utterances were used as reference selected using the casual training method.

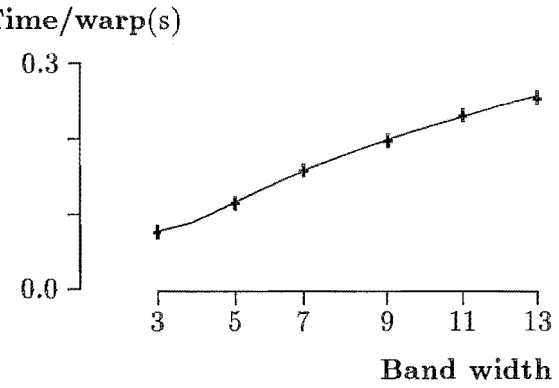


Figure 7.7. The time required for recognition with respect to the band width of a UEB method of DTW. Times are calculated for a 10 vector feature (cepstral coefficients) using Euclidean distance. Although the relative timing here is important absolute figures are given also. The figures are for calculation with a TMS320C30 non-optimally programmed in C-language.

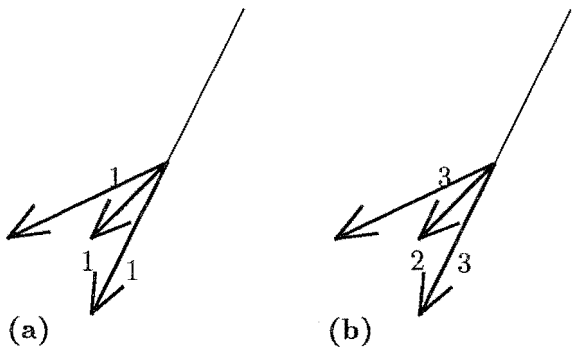


Figure 7.8. Two sets of local weights are used on the warping path of the UE21 DTW method. The path weightings affect the movement of the warping path; weightings (a) weight all paths equally, this weighting allows for equal movements in all allowable directions (constant constraints); weightings (b) weight paths unequally and thereby reduce path movements along the longer paths (non-constant constraints).

tion accuracy. Two different sets of local weightings on the local distance constraints were tested for the UE21 method. The two local weights are shown in Fig. 7.8.

For the first test, equal weightings were given to all paths (shown in Fig. 7.8(a)) and is here referred to as *constant weighting.* For the second test, weighting proportional to the sum of the warping path movement in the x and y directions was used (shown in Fig. 7.8(b)) and is here referred to as *non-constant weighting* (refer §5.1.2.4). The weightings affect the movement of the warping path so that if greater weightings are placed on the longer path movements these movements are less favoured. Preliminary tests were undertaken for one male and one female speaker only. Twenty representations of each of the ten words in the vocabulary ZERO through NINE were used with ten representations per word forming reference templates. The features tested were RMS, ZX, LPC, CEP and ZCEP (refer Chapter 4). The results for each weighting type and for each feature tested are given in Table 7.2. Accuracy improvement from the constant weighting type to the non-constant weightings type was significant (greater than 95%

level based on a paired-t test).

Feature	Distance	Gender	Constant Weightings	Non-Constant Weightings.
			Accuracies%	
RMS	Euclid	male	48.0	55.9
		female	42.8	53.5
ZX	Euclid	male	63.5	68.5
		female	43.0	60.8
LPC	Euclid	male	59.1	83.3
		female	74.6	89.3
CEP	Euclid	male	73.8	88.1
		female	88.7	93.3
ZCEP	Euclid	male	84.0	88.2
		female	86.3	94.3
Average	Euclid		66.4	77.5

Table 7.2. Accuracies for UE21 DTW method with two types of local weightings. Recognition accuracies for a male and a female speaker, speaker dependent, vocabulary consisted of the ten words ZERO through NINE with ten utterance per word used as reference and selected using the casual training method.

Another set of weightings was tested for the UEB method. The weightings were such as to weight the warping path more heavily when it deviated from a linear path. It was found that if the weightings were made too heavy the path did not deviate from the linear model. If they were made too light the path would deviate too greatly. In either case the recognition accuracy was less than for other DTW methods. After a number of weightings were tested the weighting for the band method was finally set to

$$Weighting_{UEB} = 1 + ABS\left[m - \left(\frac{N}{M}\right)n\right] \tag{7.1}$$

where m is the movement in frames of the warping path along the m-axis direction, with M being the length of the reference word and n is the movement in frames of the warping path along the n-axis direction, with N being the length of the test word (refer §5.1).

7.6.3 Summary

In the tests reported in §7.6 the UE21 method gave significantly higher accuracies than the CE21 method, hence the former was chosen for the isolated word recognition algorithm. From results reported in §7.4.2 the non-constant weightings were considered optimum for the isolated word recognition algorithm although tests with non-constant and constant weightings were undertaken. For the continuous recognition scheme, reported in Chapter 9, a UEB method was chosen with weighting applied as discussed in §7.4.2.

7.7 SEARCH TECHNIQUES

To speed up the operation of a DTW procedure many search techniques have been proposed. It has been claimed that incorporating search techniques into DTW procedures can halve the calculation time without reducing the accuracy (Bisiani and Waibel, 1982; Rabiner and Levinson, 1981).

A range of methods were first discussed in §5.1.6. Three of these methods, the branch-and-bound, the beam-search and the threshold, appeared highly successful without requiring much alteration to the DTW algorithm. However due to their individual limitations, as discussed in the following sections, a hybrid scheme is developed that incorporates the best points of these algorithms. This hybrid scheme is discussed in §7.7.3

7.7.1 Branch-and-Bound (BB) and Beam-search (BEAM) techniques

The branch-and-bound (BB) and beam-search (BEAM) techniques, first discussed in §5.1.6, are useful for their ability to speed up the DTW operation. These methods were originally proposed by Bisiani and Waibel(1982). Both these methods yield considerable computational savings ($> 60\%$) with no reduction in accuracy. Computational savings occur because these schemes terminate DTW operations for a particular template when the accumulated distance, calculated along the warping path, exceeds a threshold bound. No accuracy is lost with these methods, however, because the threshold is constantly updated during the recognition process. The threshold can be updated as warping proceeds because all the reference templates are warped in parallel. Thus during recognition, the current threshold is set based on a multiple of the minimum warping distance of any of the simultaneously warped words. At the end of computation, after all the warping paths which produce distance larger than the threshold have been terminated, the warping path with the minimum distance is always obtained. Bisiani and Waibel(1982) call the warping of all the reference templates in parallel a *breadth-first* DTW method (rather than a *depth-first* DTW method, which is the usual case).

The branch-and-bound and the beam-search techniques have the advantage of not requiring any alterations to the DTW algorithm. However, they do require a large management overhead because all the warping paths must be monitored for path termination and stored simultaneously. This requires a large memory space to store multiple sets of data. Since real-time systems typically have limited memory, these search methods are not useful for real-time applications.

7.7.2 Threshold Method

Rabiner and Levinson(1981) discuss another method of reducing the computation of a DTW algorithm. Their method uses a rejection threshold which sets an upper bound to the calculated accumulated distance of the DTW algorithm (refer §5.1.3). Accumulated distance calculations that are larger than this rejection threshold cause the DTW algorithm to terminate calculations for that reference word and to begin on another reference word for the test. The rejection threshold is set prior to DTW calculations and its magnitude is set relative to the frame number of the calculation by setting a rejection threshold minimum (T_{min}) and a rejection threshold slope (T_{slope}), viz

$$\text{Rejection Threshold} = T_{min} + (\text{Frame number})T_{slope}. \tag{7.2}$$

The variation of this threshold with frame number is illustrated in Fig. 7.9.

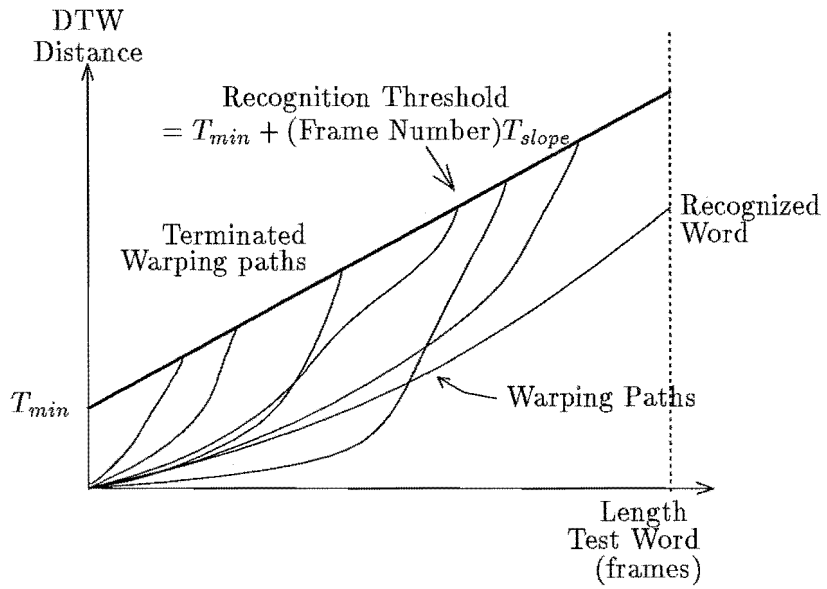


Figure 7.9. Illustration of threshold technique proposed by Rabiner and Levinson(1981). The rejection threshold cuts off warping paths when distances calculated for these paths exceed threshold limits. The threshold value is dependent on T_{min}, where the line crosses the distance axis, and T_{slope}, the slope of the line.

Because the rejection threshold function parameters (T_{min} and T_{slope}) must be set prior to calculation of the distances, the range of the distances must also be known prior to DTW calculations. This will not be a problem for a practical system where the statistics of the system are well known. However, if the system, or the surrounding noise environment changes, the distances calculated may change radically. As shown in §6.2, for cepstral coefficients, an increase in noise levels can drastically reduce the norm and variance of the norm, affecting the distance values calculated. To allow for such changes, and to ensure that a final recognition decision is made, a distance threshold must be set so that at least one DTW distance is smaller than the rejection threshold. By increasing or decreasing the threshold to allow for any changes, as discussed above, problems with this method can occur. These problems can be caused by either increasing the threshold which allows larger numbers of computations defeating the purpose of using the rejection threshold, or decreasing the threshold and hence allowing only a few templates to fall below the threshold and to increase the number of times no word is recognised.

Another problem with the Rabiner and Levinson(1981) method occurs if feature types are changed, such as might occur with a test system. In this case thresholds must be recomputed and reset while other methods, such as the breadth first techniques, automatically adapt. The advantages of the technique, however, is that it does not require parallel, breadth-first, reference recognition, thus facilitating real-time application with limited memory.

7.7.3 Hybrid Method

A search technique which is composed of the best points of both the methods described above is reported in this section. A depth-first DTW procedure is used which uses a continually updated rejection threshold for DTW termination. This rejection threshold

is updated based on previous recognition calculations (and based on the accumulated DTW distance measure, refer §5.1.3) rather than set to an absolute value prior to recognition. Because the threshold is based on the accumulated distance measure the threshold is known as the *accumulated rejection threshold.*

At the start of DTW calculations the accumulated rejection threshold is set to ∞. The first accumulated DTW distance is calculated with a reference template. Note that the accumulated distance is normalised with respect to the length of the warping path and hence is equivalent to the average local distance along the warping path. The accumulated distance threshold is found by scaling the accumulated distance to a multiple of its level. Subsequent local distance calculations, which are normalised during calculation and so are equivalent to the averaged distances calculated along the warping path, are compared to this accumulated distance threshold. If a local distance calculation is larger than the accumulated distance threshold calculations are terminated. If the local distance values are less than, or equal to the accumulated distance threshold DTW continues until an accumulated DTW distance is calculated and becomes the minimum accumulated distance. At the end of a DTW calculation that has not been terminated the accumulated rejection threshold is updated to become a multiple of this minimum accumulated distance. Hence the threshold continually shifts to a multiple of the minimum accumulated distance calculated during the recognition procedure.

By terminating recognition with respect to a multiple of the minimum accumulated distance, time for recognition calculations is reduced by up to 50%, while recognition accuracy is not affected. Different multiples of the minimum accumulated distance (which is equivalent to the accumulated rejection threshold) were tested. Values tested caused termination when the accumulated rejection threshold was 200%, 160%, 120% and 100% of the minimum accumulated distance. Table 7.3 shows the calculation time versus the accuracy for these different levels. Recognition times are given for a feature vector of ten dimensions (in this case ten cepstral coefficients) and averaged over experiments using 2, 6 and 10 reference templates. Word lengths were, on average, 35 frames long and each frame, containing the set of ten coefficients, is derived from 200 samples and sampled at 10kHz. A Euclidean distance measure was used.

Minimum Accumulated Distance multiple	Accuracy(%)	Average time(s)
∞	84.6	1.027
2	84.6	0.754
1.6	84.6	0.520
1.2	80.17	0.381
1.0	76.3	0.343

Table 7.3. Accuracies versus speed of DTW calculation with respect to accumulated rejection threshold value. The accumulated rejection threshold value is set as a multiple of the minimum accumulated distance.

7.8 SUMMARY

The tests discussed in this chapter have been formulated to reduce the variability in choice of many of the requirements of the recognition system. Variables that have been discussed in this chapter and that will be set to constants (for the tests discussed in Chapter 8) include the RMS level for endpointing, the methods used to extract the features that are tested, the search technique to speed up recognition, and the DTW constraints. Selections which are reduced have included the recognition training methods (clustering or random selection), and the DTW method (CE21, UE21, or UEB). The selection of these variables were based on the general use of these features and methods in the literature, the ability for recognition (calculated as the averaged accuracy over the features and speakers to be tested), and also the capability of the system operating in real-time. Having established methods of endpointing, pre-processing, acoustic feature extraction and dynamic time warping algorithm, these methods are incorporated into the structure of a flexible isolated-word recognition system as shown in Fig. 7.1. The recognition scheme is used to compare, in Chapter 8, the abilities of acoustic features, pre-processing techniques and distance measures for the recognition of isolated words.

Chapter 8

ASSESSMENT OF RECOGNITION FEATURES AND PROCESSING EFFECTS

In this chapter a number of pre-processing variables and recognition features (refer §7.4 and Chapter 4) that are useful for word recognition algorithms are assessed. The factors examined in this chapter can severely affect the outcome of a recognition system. The particular factors assessed are pre-processing variables, feature types, speaker variation, and distance measures. Other factors that also affect recognition systems, such as vocabulary, noise and database parameters (sampling, recording effects, etc) are kept constant for all tests.

During the testing of the recognition systems it has been found that it is important to consider a wide range of factors when comparing results from different recognition schemes. Lea(1982) identified over 80 factors, including task factors (type of hardware involved), human factors (speaker-dependent factors), language factors (how words are spoken and the type of language used), channel and environmental factors (noise, distortion and effects on the speaker), algorithmic factors (the exact form of the algorithm implemented), performance factors and response factors (how the algorithms respond to errors). These factors affect many parameters by changing the time and frequency structure of the speech. If the relative accuracies of different recognition devices are to be compared, all of these factors must be known for each device. To be able to make an informed decision on the merits of a recognition device it is essential to know the relative effects any changes in these factors have on recognition accuracy. The effect of changing factors such as the vocabulary, the prosodic nature of the voice, and the recognition system, is little understood and difficult to quantify. In such cases it is difficult to predict a device's recognition accuracy.

It is, however, possible to determine the relative affect of particular variables on a recogniser's accuracy. To do this each variable must be carefully tested while controlling all other areas of possible variation within the recognition system (providing that the variables are independent). Recognition accuracies can then be carefully noted along with all system parameters thus obtaining the effect on recognition accuracy of altering that particular variable.

8.1 STANDARD RECOGNITION FACTORS

In order to carry out objective tests of each recognition factor the recognition algorithm and testing conditions are standardised. The following sections discuss those factors which are held constant during testing.

8.1.1 Vocabulary

A standard vocabulary of the words ZERO to NINE is used for all tests. This vocabulary is chosen because an American speaker database, containing these same words,

is available and hence comparative studies with New Zealand speakers can be performed allowing the comparison of recognition performance for speakers with different accents. Another reason why digits are often chosen is because the recognition of digits lends itself to useful and interesting applications of speech recognition (Rabiner and Sambur, 1976). Digit recognition allows applications such as telephony, voice input to machine, banking by voice, and inventory and stock record keeping by machine.

The digit vocabulary is one of the more difficult to recognize (Rabiner *et al.*, 1986) since it consists of single syllables and highly confusable words. Rabiner(1986) equates the difficulty of recognizing the digit vocabulary, based on recognition accuracies, with a 54 word computer term vocabulary and a 96 word American states vocabulary (a vocabulary of the name of each state in America). Although there are only ten words in the digit vocabulary it has a good mix of voiced and unvoiced sounds, containing vowels, stops, plosives, unvoiced fricatives and voiced fricatives.

This same vocabulary has been used many times since its first use in the early 1950s by Davis *et al*(1952). It has been used as a bench mark vocabulary for recognition of isolated words, (Denes and Mathews, 1960; Olson and Belar, 1960; Teacher *et al.*, 1967; Gilli and Meo, 1967/68; Purton, 1968; Ewing and Taylor, 1969; Scarr, 1970; VonKellar, 1971; Warren, 1971; Ichikawa *et al.*, 1973; Sambur and Rabiner, 1975), and connected utterances, (Sambur and Rabiner, 1976; Rabiner and Sambur, 1976; Rabiner and Schmidt, 1980; Rabiner *et al.*, 1984b; Rabiner *et al.*, 1989). It has also been used in other languages (Lau and Chan, 1985), and combined in other word sets such as the alphabet (White and Neely, 1976; Myers *et al.*, 1980; Das, 1982) and airline words (Rabiner *et al.*, 1984b). By using the same vocabulary, researchers have compared systems and stated relative abilities. This comparison is, however, only accurate in as much as all the other recognition factors between the systems are kept constant.

8.1.2 The Word Matching Algorithm

Recall from §5.1 and in §7.6 that a dynamic time warping algorithm is chosen as the recognition method. This method is chosen for a several reasons. Firstly, training is easier than for other approaches, such as hidden Markov models and neural nets. This is because, for these other recognition methods, recognition accuracy can be reduced if large numbers of reference templates are not used to produce appropriate word models. In the case of hidden Markov models up to 1000 reference word examples are required in the training process (Rabiner and Levinson, 1981). For the dynamic warping method as few as 10 templates for each desired word need only be saved and the recognition system run. Secondly, because dynamic programming systems have been used for many years their recognition ability is well known. Thirdly, for the dynamic time warping method (as well as other methods such as HMM), we suppose that the accuracy of the features tested is not compromised by the recognition scheme. This is important if factors are to be compared and assertions made on the relative abilities of these factors. If this assumption is not valid, that is, if it is assumed that each feature interacts with the recognition algorithm, then the relative abilities of the features cannot be made.

8.1.3 The Database

Two databases are employed for the recognition experiments described in this chapter. The first consists of ten New Zealand speakers, (6 male and 4 female), each speaking the digits, ZERO through NINE, up to 20 times. The second is of 20 American speakers (11 males, 9 females) each speaking the digits, ZERO through NINE, twice.

The New Zealand speakers were recorded in the Electrical Engineering laboratories of the University of Canterbury. The speaker used for these experiments are those also

used by Elder,1991 , and they are referred to individually by a two letter abbreviation of their name (eg AM, AE, CW, etc). Two sets of recordings were made. In the first recording five of the ten speakers were recorded and in the second recording, one year later, the final five speakers were recorded. The speakers uttered the words ZERO through NINE in sequence once every morning and again in the afternoon. The participants were asked to speak the words as naturally as possible pausing slightly between words. All recordings were carried out in the same room to keep background room conditions such as noise, echo, and reverberation constant from recording to recording. The words were recorded using a handheld microphone which was held by the participants at a comfortable distance from the mouth. This distance was not monitored and is one variable which may introduce recognition errors (Kahn and Gnanadesikan, 1986; Baker and Pinto, 1986). A different microphone were used for the second set of recordings than that used for the first set due to the time span between recordings. Both microphones had similar performances. The first was an AIWA CM-53 microphone which has a flat frequency response from 50 to 13000 Hz. The second microphone was an Audio-Technica AT818II which has a flat frequency response from 50 to 15000 Hz. The sentences were recorded by an AIWA F990 cassette recorder on to a low-noise tape employing Dolby C noise reduction. The transfer function of the tape was determined to be flat between 20 and 18000 Hz and the phase linear (to within 10°) between 100 and 5000 Hz (Thorpe, 1990). The words were digitised using an SX10 digital audio board manufactured by Antex Electronics (Antex Electronics Corporation, 1990) and the digitised utterances stored on an IBM-PC hard disk. The speech was later interactively inspected by examining plots and listening to audio output for any irregularities. The speech was digitised at 10kHz after being filtered by a 4.4kHz seven pole elliptical low pass anti-aliasing filter. The magnitude and phase response of the elliptical filter are shown in Fig. 8.1.

The words were spoken as connected word sentences and were later manually segmented into individual words. Endpoints of the segments were chosen to leave at least 15ms of silence at the beginning and end of each word, where possible. The individual words were stored back on IBM-PC hard disk for subsequent use.

Speaking intensity often changed during an utterance, both increasing and decreasing the utterance intensity. This is illustrated in Fig. 8.2. Some utterances contained noise spikes from microphone knocks during recording. Also, breath noises were often made by speakers during (usually in the middle) and at the end of speaking. Sometimes words were run together, particularly the words FIVE-SIX-SEVEN. Examples of these recording aberrations are shown in Fig. 8.3. All of these speech segments were used in the recognition testing because it was considered important to have naturally varying speech sounds obtained from normal speaking conditions.

The other database used in the experiments is an American speaker database, purchased from the National Institute of Standards and Technology (NIST). This data was digitised at 20kHz with a 10kHz anti-aliasing filter. It was assumed, because of the 10kHz cutoff of the aliasing filter, that the aliasing filter had flat frequency and linear phase responses between 50 and 4500 Hz, the frequency range of interest in this study. The utterances of the database were copied from CD-ROM to IBM-PC hard disk. In order to compare results using the American speakers' data with results using the New Zealand speakers' data, the USA data was decimated to reduce the sampling rate to 10kHz and then filtered with a software simulation of the anti-aliasing filter used on the New Zealand data. To simulate the anti-aliasing filter I choose to use either a software simulated 8 pole elliptical filter approximating both the magnitude and phase of the filter used when sampling the NZ speakers speech (an 8 pole simulated filter was needed to accurately model the 7 pole elliptical filter within 1% ripple variation in the

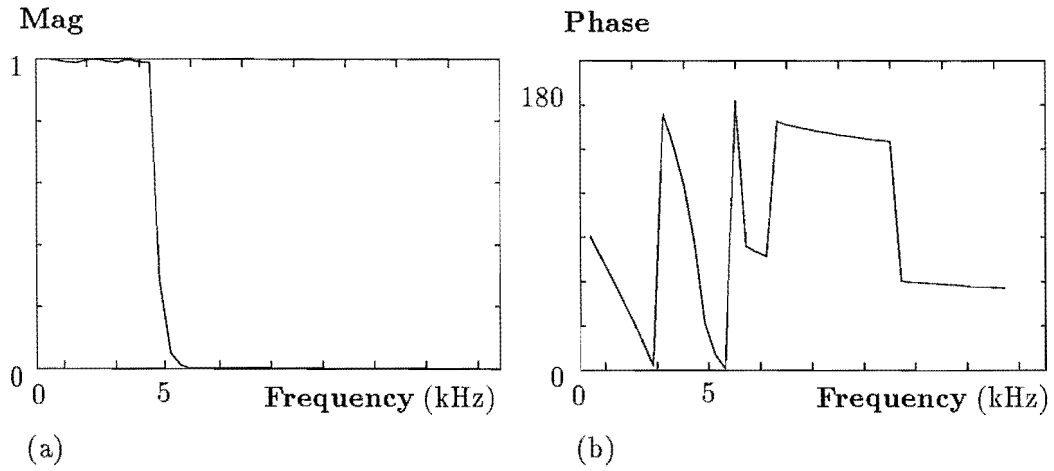


Figure 8.1. Magnitude (a) and phase response (b) in the frequency domain of the 7 pole elliptical filter used as an anti-aliasing filter before digitising the data.

passband), or an eight point FIR filter approximating only the magnitude of the filter used when sampling the NZ speakers speech. The advantage of the FIR filter was that it is simpler than the elliptical filter and, because only 8 taps were required to simulate the magnitude response within 1% ripple variation, it took less time to operate.

To compare results between the NZ and USA databases it was important to ascertain whether differences in recognition accuracy occurred due to differing filter types. Thus preliminary recognition tests were undertaken to find if it was necessary to simulate both filter magnitude and phase responses for the American data to give the same filter response as that used for the sampling of the New Zealand data. Recognition tests were performed on the American data using low-pass FIR and elliptical filters. Both of these filters had the magnitude response, depicted in Fig. 8.1(a), but different phase responses, the FIR filter having linear phase and the elliptical filter having the non-linear phase response pictured in Fig. 8.1(b). Recognition tests were undertaken to find any differences in recognition accuracy due to phase effects of the filter devices. For the recognition tests two reference templates per word were stored and two tests per word were performed. The tests were jackknifed (Mosteller, 1971), rotating the test and reference words, so that each test word was used as a reference word and each reference word was used as a test word. This rotation method allowed a total of four unique tests before the combinations were repeated. The jackknife procedure and how it was implemented in this thesis is more fully discussed in §8.1.5.1.

Linear prediction coefficients (LPC), cepstral coefficients (CEP), and root-mean-squared (RMS) features (refer Chapter 4) were tested. Error rates between the two types of filtered data were only slightly different, although the particular word in error was often different. One such test produced identical error rates but 50% of the errors between the two systems were with different words. It was obvious from these tests that the phase response of the filter does affect the accuracy of the recogniser using these feature types. Although it may seem obvious that the phase distortion of the filter will affect the magnitude of the speech and hence change its RMS it was not known why it should affect the frequency information of the speech which is represented by the LPC and CEP features. Because these features were affected it was therefore decided to continue applying the elliptical filter to the data of the American database hence simulating the magnitude and phase of the filter used for the New Zealand database.

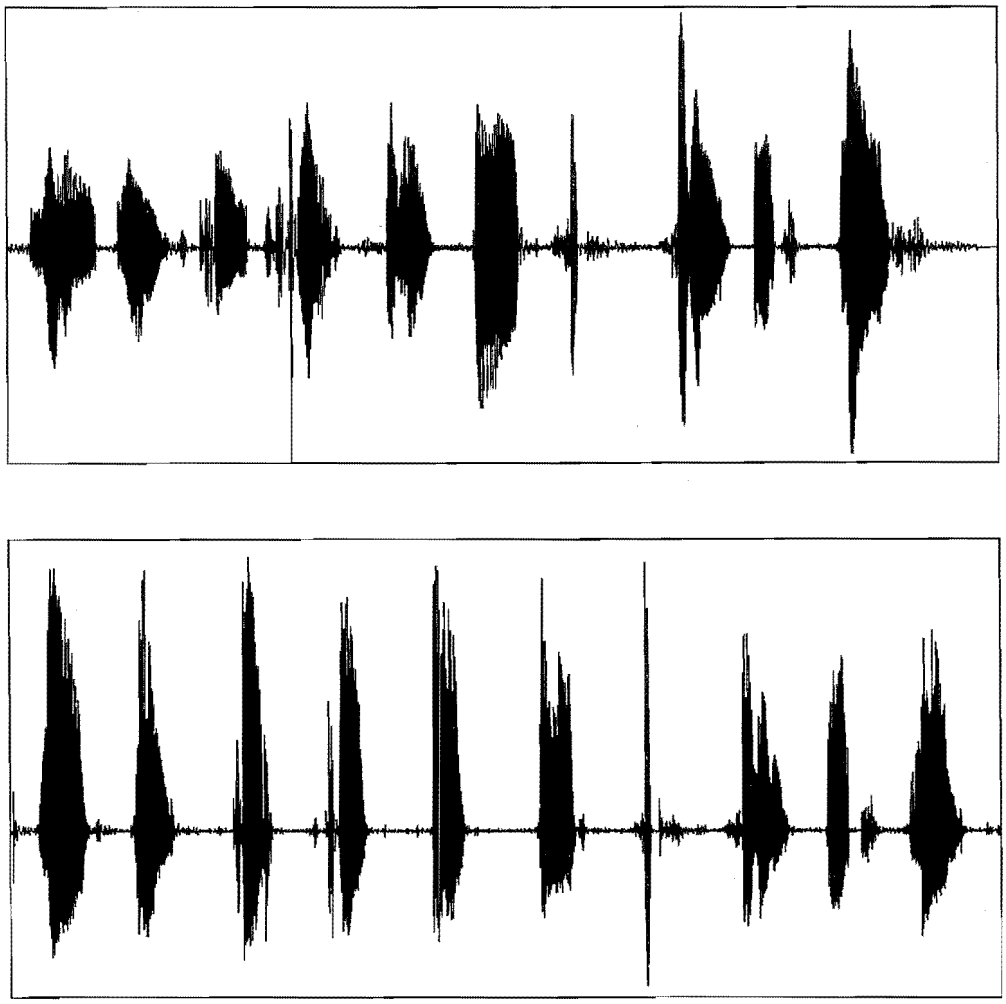


Figure 8.2. An example of the loudness variations that occur during a sentence recording: (a) the speech waveform of a speaker showing increased utterance intensity (JK1), (b) the speech waveform of a speaker showing decreased utterance intensity (CP15).

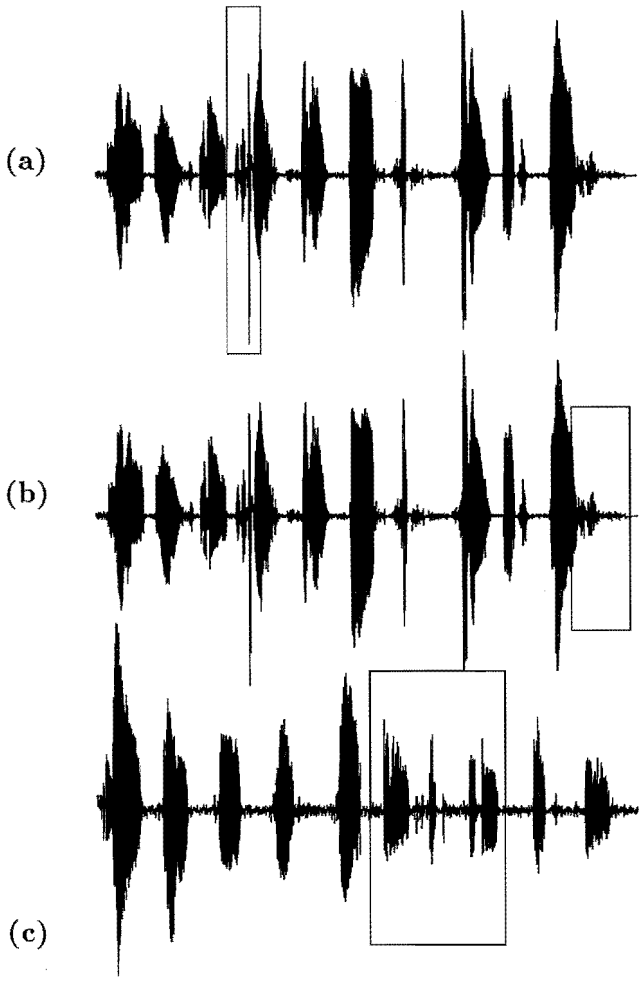


Figure 8.3. Example of some of the recording aberrations. (a) Noise spikes, usually from microphone knocks(CP15). (b) Breath noises introduced between and at the ends of utterances (CP15). (c) Lack of pausing between utterances (AW1).

8.1.4 Noise Level of the Database

An average noise level for the New Zealand data was calculated from the recorded sentences. The mean signal to noise level was found to be 29dB with a 14dB variance. Large variance occurred because some sentences were recorded with very little noise while other sentences had audible background noise. The signal to noise ratio was calculated as

$$\frac{S}{N} = 10 \log_{10} \frac{V_{s+n} - V_n}{V_n} \tag{8.1}$$

where V_{s+n} is the variance of the signal taken during the voiced periods of the data and V_n is the variance of the signal taken during the silence periods. The voiced and silence parts of the speech were found by examining the speech and separating the regions from the rest of the speech by hand.

The American data was recorded in an acoustically treated sound room (Tracoustics, Inc., Model RE-244B acoustic enclosure) with the microphone placed 2-4 inches

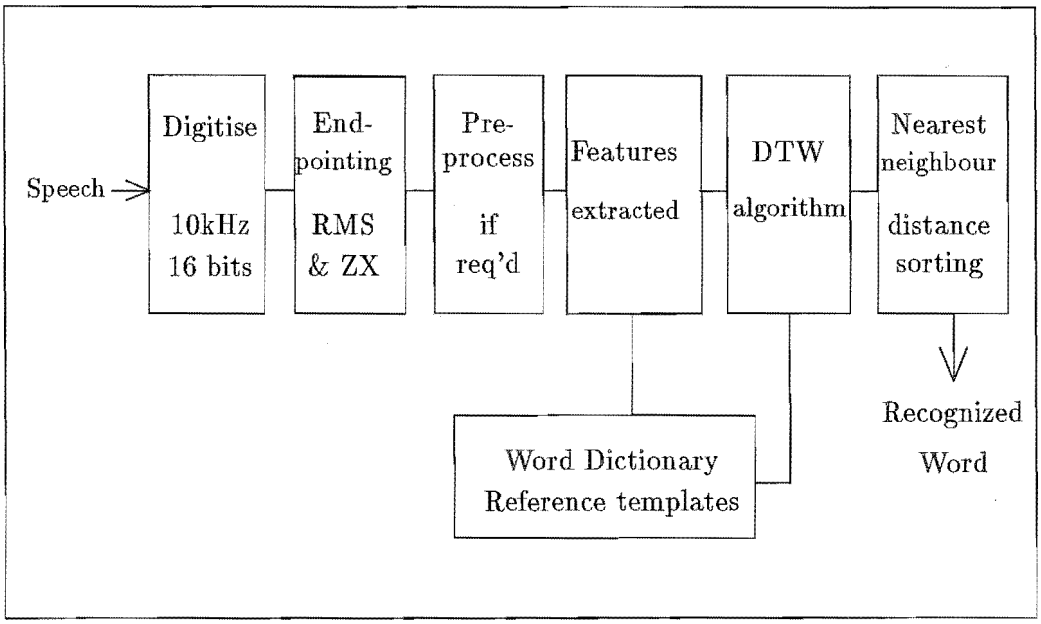


Figure 8.4. Block diagram of real-time recognition algorithm used to test varying parameters.

in front of the speaker's mouth. The microphone was an Electro-Voice RE-16 Dynamic Cardioid. No other data was available for this data base such as filter type, filter phase and magnitude response, noise level, recording times and frequency of recording per person.

8.1.5 Testing conditions

A block diagram of the real-time system used to test various recognition parameters and features is shown in Fig. 8.4. Details of the system tested in this chapter, such as the endpointing, feature extraction methods, types of DTW, methods of training, and distance measures are described in Chapters 5,6, and 7.

A paired-t test is used to test statistical significance for the results quoted. Significance was set at a confidence level of 95% The tests were carried out on each of the New Zealand speakers using a speaker-dependent system. Recognition accuracies are given for ten reference templates per test word unless otherwise stated. For all of the following recognition experiments a UE21 DTW (refer §5.1 and §7.6) method was used either with constant weightings or non-constant weightings on local constraints (refer §5.1.2.4).

8.1.5.1 The implementation of the Jackknife procedure

The jackknife procedure (Mosteller, 1971) allows multiple trials to be run on the same dataset. The jackknife procedure exchanges the test and reference templates so, for each new trial, the reference and test data sets are different. The trials are used to calculated the mean and standard deviation. The jackknife procedure has two purposes firstly, to eliminate bias of the order of 1/(sample size) within the biased sample statistic and secondly to give an honest measure of variability based upon the data especially for complicated problems with limited data (Mosteller, 1971). Usually statistics are calculated from a group of data by breaking the data into sub-groups and calculating the variability from one group to another. Sometimes, such as the case of recognition

experiments, the statistics require so many observations that it cannot be calculated from the sub-groups even if there are only two sub-groups, in such cases the jackknife method offers a way out. Although the jackknife method uses sub-groups, it gets its variability from differences between the statistic computed on all the data but varying the sub-groups with this data (this is often achieved by removing some of the data from the sub-group). By computing statistics with each sub-group we get a set of numbers that can be used as normally distributed data, leading to t-statistics.

For the recognition test performed thoughout this thesis the test and reference datasets of words were chosen randomly from the original dataset. Generally this was achieved (where all 20 templates were used) by randomly choosing 10 templates for the test and using the other 10 templates for the reference data. In most of the trials discussed the tests and reference templates were simply rotated. For example, for the first recognition test the reference data templates are numbered $r_0, r_1, ..r_9$ and the test templates are numbered $t_0, t_1, ..t_9$ then for the second set of tests the reference template set would consist of $t_0, r_0, r_1..r_8$ and the test templates data set would consist of $t_1, t_2..t_9, r_9$. Thus, for each trial, one test and one reference template are exchanged producing different test and reference data sets. This rotation method allowed twenty tests from a reference and test set of ten templates each. Thus 200 trials (20 tests for ten speakers) allowed 2000 recognition tests to be performed (because each speaker had 10 test words). Generally a full rotation of 20 different trials was performed. However, due to the time consuming nature of this number of tests, for some recognition testing fewer rotations were used. Wherever the testing procedure differed the actual use of the jackknife is discussed in the text along with the testing procedure.

8.2 TESTING OF RECOGNITION PARAMETERS

The following sections give results from testing various recognition parameters. Generally, these results have been given (as graphs) as average values, either with respect to the speaker or feature tested. Those individual features or speaker which have not followed the general trend have been noted specifically in the text. The averaging of results has been done to show general trends when applying certain processing techniques to the speech because the absolute values of each test is also dependent on the speaker dependent parameters such as loudness, pitch, speaking rate, accent etc. It was hoped that by averaging these results these speaker dependent effects may also be averaged out and hence show those trends more for the parameters being tested. Where average plots have been given, the standard deviation of these plots have also been shown as a vertical bar drawn through the center of each bar graph.

8.2.1 Pre-processing

Recognition algorithms use pre-processing such as filtering, pre emphasis, windowing and window overlap as standard techniques. All these different techniques are considered under the heading 'pre-processing' to distinguish them as types of processing that can be varied prior to feature calculation for all features (refer Chapter 4). These pre-processing steps are considered important to produce accurate speech representations (as for speech synthesis). However, the importance of each of these processes for recognition along with how much these processes affect the recognition accuracy is unknown. A precise representation of the speech information, involving speaker-dependent traits as required for high quality speech synthesis, may not be important for speech recognition. In fact some researchers claim higher recognition accuracies when less precise information is retained, (Hermansky and Junqua, 1988; Niederjohn *et al.*, 1987).

Recognition accuracies with and without pre-processing techniques were examined to ascertain their importance. The pre-processing techniques examined were those usually found in recognition schemes, notably pre-emphasis, type of window, overlapping speech frames and frame size.

Knowledge of the influence on recognition accuracy of these techniques is important because it may allow a reduction of the computational overheads in practical recognition schemes. The effect of some variables such as frame size can not only affect the accuracy but can also affect the speed of operation, since the computations required for dynamic programming are proportional to the square of the number of frames of the speech (and the larger the frame size the fewer the number of frames required computation). Another variable which affects the computational requirements is the number of reference templates stored per word. If fewer templates are used the number of dynamic time warping operations can be substantially reduced.

8.2.1.1 Frame Size

Frame sizes of 100, 200, 300 and 400 samples were tested. The accuracies for each frame size, averaged over all features and pre-processing effects, is shown in Fig. 8.5. The accuracies with respect to the gender of the speakers are also shown.

By examining the data of Fig. 8.5, for both the male and female speakers, using a paired-t test, it is found that there is a significant difference in the mean accuracies between frame sizes of 200 and 400 samples, with higher accuracies obtained for the 200 sample frames. 200 sample frames also had slightly higher accuracies than 100 and 300 sample frames, although a paired-t test did not show this to be significant. Examining the speakers by sex, the female speakers show no significant difference in accuracies between frames of 200 and 300 samples, but show a significant accuracy increase when the frame size is reduced to 100 samples. For male speakers the accuracy for frames of 200 samples is significantly higher than for frames of 100, 300 and 400 samples. These results suggest that accuracy is related to the pitch of the voice, since highest accuracies occurred with the frame size approximately two to three times the speakers' pitch period.

Comparing the individual speaker's accuracies directly with their voice pitch (also shown in Fig. 8.5) shows a relationship between the pitch period and the recognition accuracy with respect to frame size. Although this relationship is somewhat tentative because of the small data set, it appears that the higher the pitch of the speaker's voice (such as for speakers CW and MC) the smaller the frame size required to obtain optimum accuracy. Low pitched voices however tend to require larger than average frame sizes (DR and JK). However, this did not hold for all speakers. For example, speakers AE and AM have roughly the same pitch but optimum performance was obtained with frames of different sizes.

Obtaining the speaker's pitch prior to word recognition is not an option because it is difficult and time consuming to calculate. For a practical system it would be reasonable to simply switch between set frame sizes for male or female speakers or choose a frame size which gives reasonable performance for all speakers (say at 200 samples).

Higher accuracies for frame sizes of around two pitch periods seem reasonable when a section of speech is examined. Extracting too short a section of speech can cause features to vary rapidly from frame to frame and hence the features do not characterise the sound well. This is caused by irregularities in the placement of the frame with respect to the pitch. Extracting too long a section of speech can cause lose of transitional information causing sounds to be averaged and short sounds lost. This sort of information is important since it distinguishes words such as ONE and NINE.

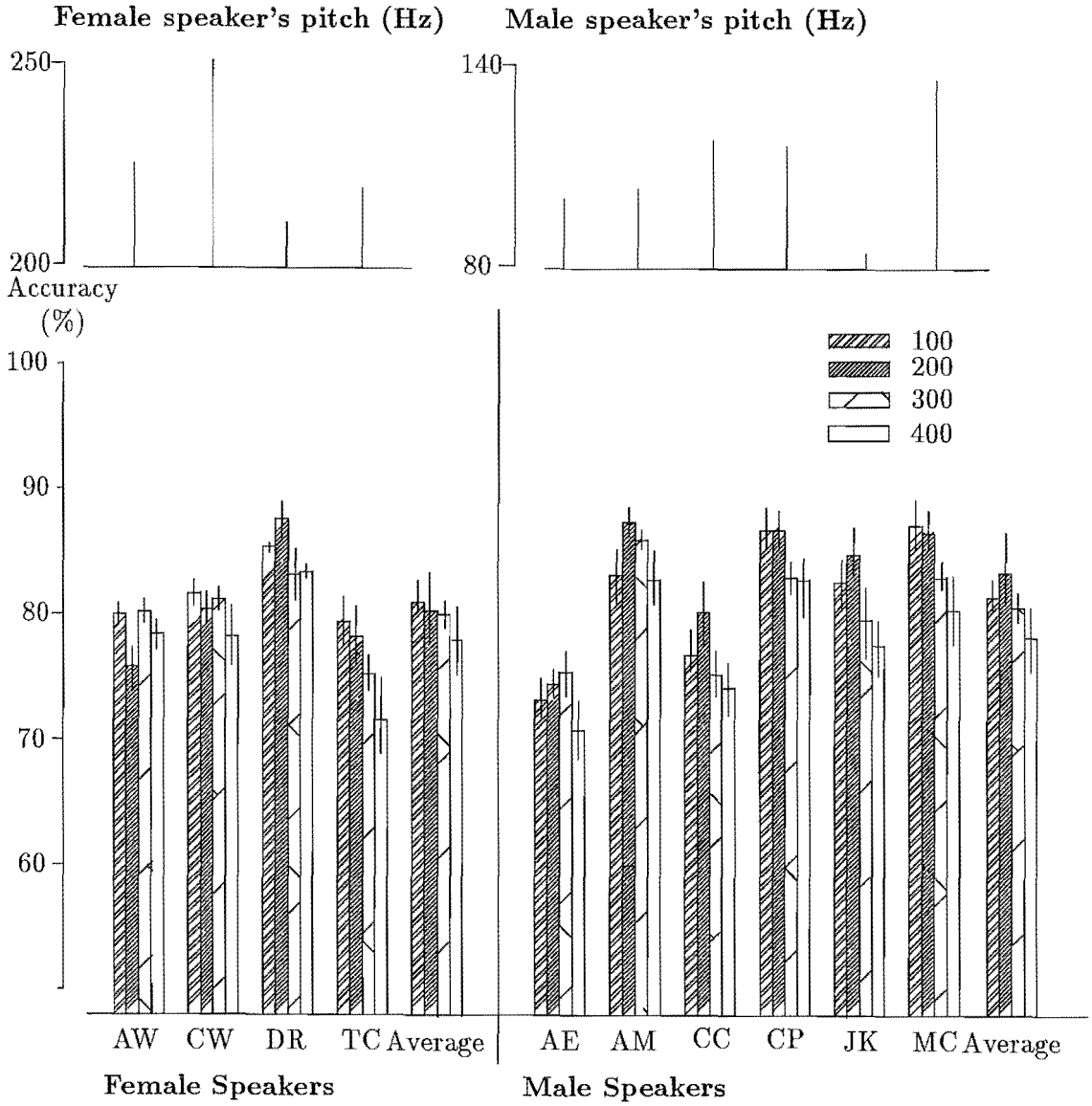


Figure 8.5. Average recognition accuracies and standard deviations (shown as a vertical line at the top of each bar) of the recognition accuracies calculated for each speaker with varying frame sizes. The average pitch of the speakers' voices is given above. Accuracies are shown for a recognition system using a UE21 DTW method with non-constant weightings. Tests performed on New Zealand speakers in speaker-dependent mode. Each test used 10 reference templates per word. A Euclidean distance is also used.

As mentioned above, for short frames of less than a pitch period, frame position (or frame placement) can seriously affect the extracted feature causing it to vary widely even though extracted from identical speech. As the frame size increases this variation with respect to frame placement is reduced. Fig. 8.6 shows an example of LPC coefficients calculated with increasing frame sizes and using a Hamming window. The graph of the LPCs show how the coefficients stabilise as the size of the window increases. One, two, three, and four pitch periods of data are extracted from a section of synthetic speech. The speech is synthesised from ten LPC coefficients extracted from the vowel sound /i/ of the word HEED. The prediction error, as defined in §4.3, is plotted in Fig. 8.7. Examining Fig. 8.7 shows that frame sizes of one and two pitch periods are

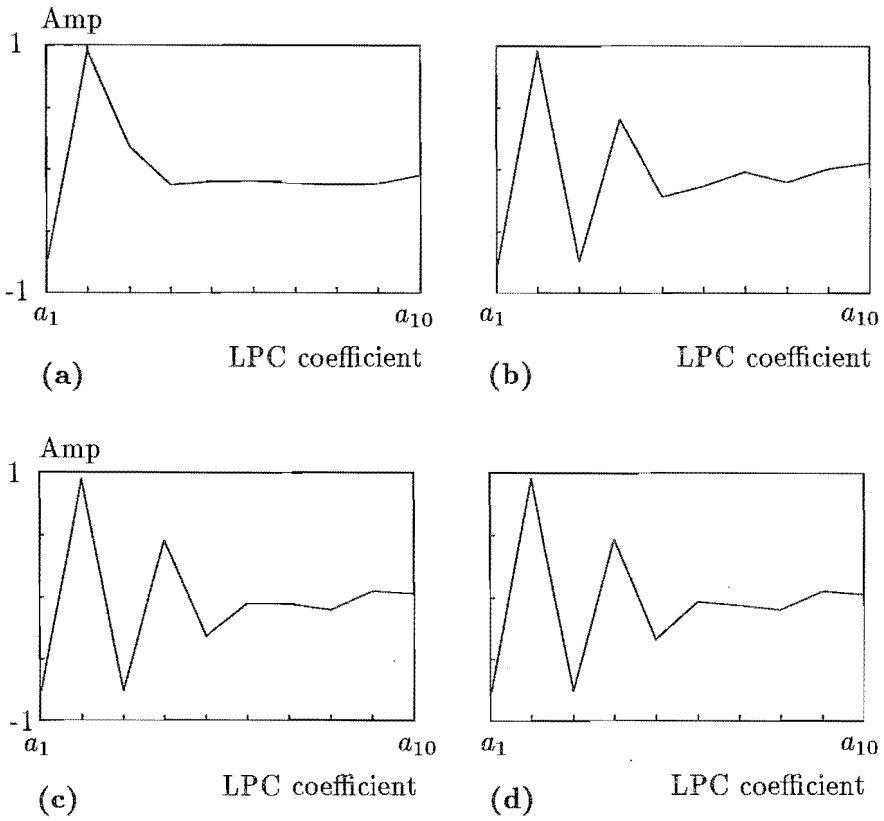


Figure 8.6. LPC coefficients extracted from synthesised speech with different window sizes for a Hamming window. (a) one pitch period, (b) two pitch periods, (c) three pitch periods, (d) four pitch periods.

relatively sensitive to frame placement. Placing a short window at pos 1 (see Fig. 8.7) drastically affects the shape of the windowed speech wave. This affect is particularly apparent for windows that alter the shape of the speech by largely reducing the amplitude of the major peak of the speech, such as a Hamming, Hanning or Blackman windows. LPCs calculated from this data are therefore altered significantly by small perturbations in the position of the window.

With larger frame sizes, distortion of the speech wave at the beginning and ending of the frame is less of a problem because the window bounds more than one pitch period. Thus, with larger frame sizes there is a lower error for all frame placements as shown in Fig. 8.7. However, even though the feature error may reduce with large frame sizes, the size of the frame may be too large to accurately represent transitions and short speech segments. If this is the case, large frame sizes may not give high recognition accuracies.

8.2.1.2 Windowing

Windowing the data with a function that smoothly reduces to (or close to) zero, such as a Hamming, Hanning or Blackman window, has been known to give more accurate feature representation than a rectangular window (Harris, 1978; Gray *et al.*, 1980; Kirkland, To be published).

In this study it was found that the choice of window type was the most critical step

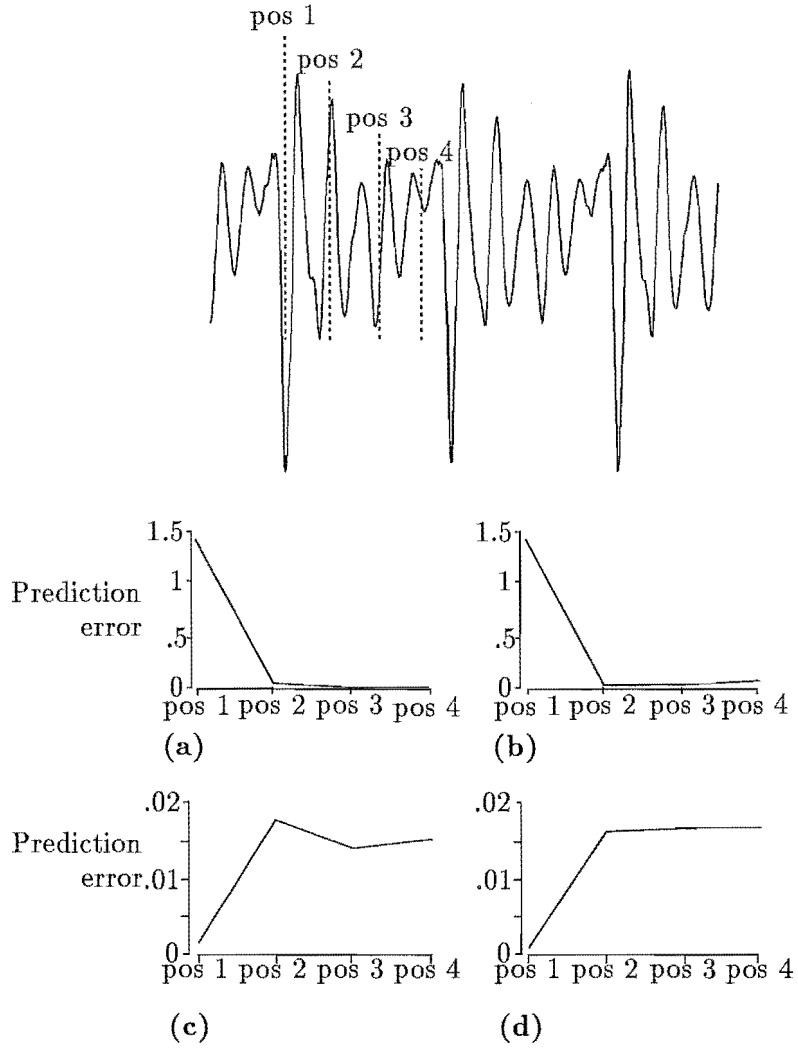


Figure 8.7. Prediction error versus frame size and frame placement showing how both frame size and frame placement can affect prediction error. Error shown with respect to window frames of length (a) one pitch period (b) two pitch periods (c) three pitch periods (d) four pitch periods. Frame placement is changed so that windows begin at one of the four positions; pos1 to pos4.

in pre-processing, from the point of view of recognition accuracy. Highest accuracies were obtained by replacing the rectangular window with one of the smoother window functions. Fig. 8.8 shows the improvement in accuracy for a Hamming window over a rectangular window. The recognition accuracy shown for each speaker is averaged over all the features tested. Averaging over all the features and windowing using a Hamming window gave a significant improvement. For both male and female speech, a Hamming window gave an average improvement in recognition accuracy of approximately 20%, (16.4% for male speakers and 25.5% for female speakers). Fig. 8.9 shows the increase in accuracy, with a Hamming window, for each individual feature. There is an improvement in accuracy for every feature except RMS. Although not shown specifically in the figure the RMS accuracy decrease, when using a Hamming window, was 15% for male speakers and 32% for female speakers. Other points to highlight (although not individually plotted) is the largest percentage increase in accuracy using a Hamming window for male speakers occurred with LPC coefficients giving an average increase of 56.4% (recognition accuracy improved from 29.8% to 46.6%), and for the female speak-

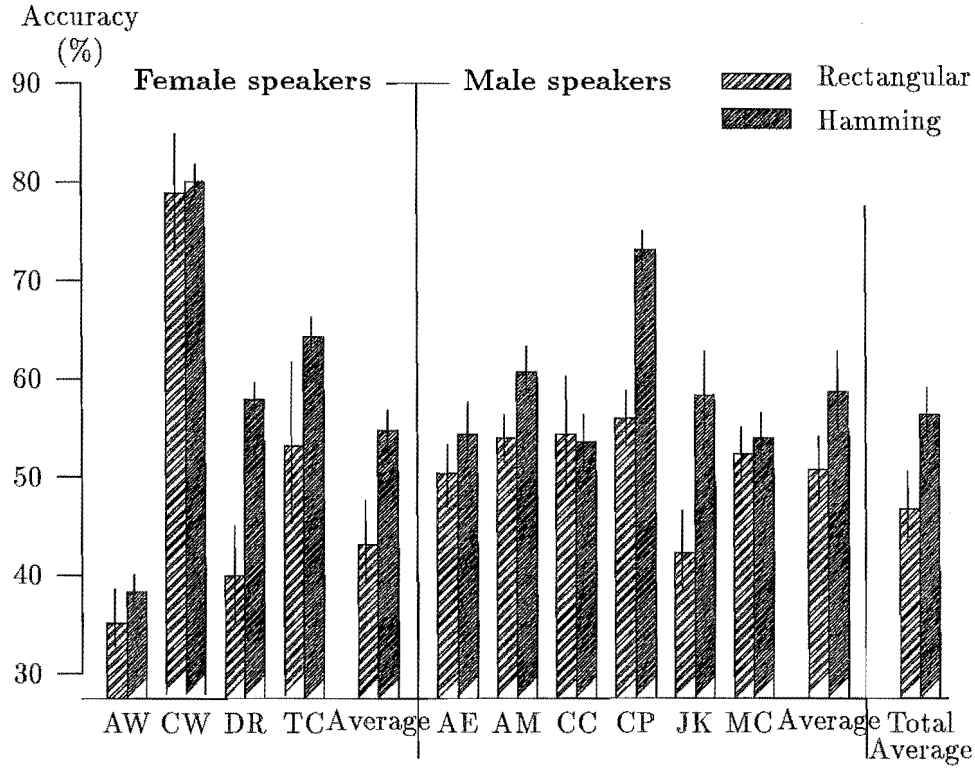


Figure 8.8. Average recognition accuracies and standard deviation (shown as a vertical line at the top of each bar) of the recognition accuracies with Hamming windowing and rectangular windowing applied to the speech data. Recognition accuracies for each speaker are averaged across all features. Results show the Hamming window to give significantly higher accuracies than the rectangular window. Results are obtained using a UE21 DTW method with constant weightings. Tests were performed on New Zealand speakers in speaker-dependent mode with 10 reference templates per word. A Euclidean distance is used.

ers, with LPC coefficients also, giving a 103% accuracy increase (recognition accuracy improved from 24.8% to 51%). For male and female speakers, cepstral coefficient also had a large accuracy increase of 43.1% and 48.8% respectively, (recognition accuracy improved from 43.4% to 62.1% for the male speakers and 44.2% to 65.8% for the female speakers).

The accuracies quoted above were averaged over all other pre-processing effects (window sizes and pre-emphasis). It was noted, however, that the accuracy improvements of the window type was tempered by the length of the window shown in Fig. 8.10. Examining the male speakers' accuracies showed an increase from a rectangular window to a Hamming window of 17.2% at 100 samples per frame, 27.8% increase at 200 samples per frame and 20.9% increase at 300 samples per frame. The greatest increase was for a frame size of approximately two pitch periods in length. The female speakers gave larger percentage accuracy increases overall, following the same trend as the male speech. Accuracy increases for the female speech were 41% for 100 samples per frame, 47% for 200 samples per frame and only 11% for 300 sample frames. These accuracy effects complement the effects discussed in §8.2.1.1 since the highest accuracy increases occurred for frame sizes of approximately two to three pitch periods. Fig. 8.11 is a plot of LPC prediction error (refer §4.3) versus window length for two window types, Hamming and rectangular, for a male speaker. From these results it is obvious that

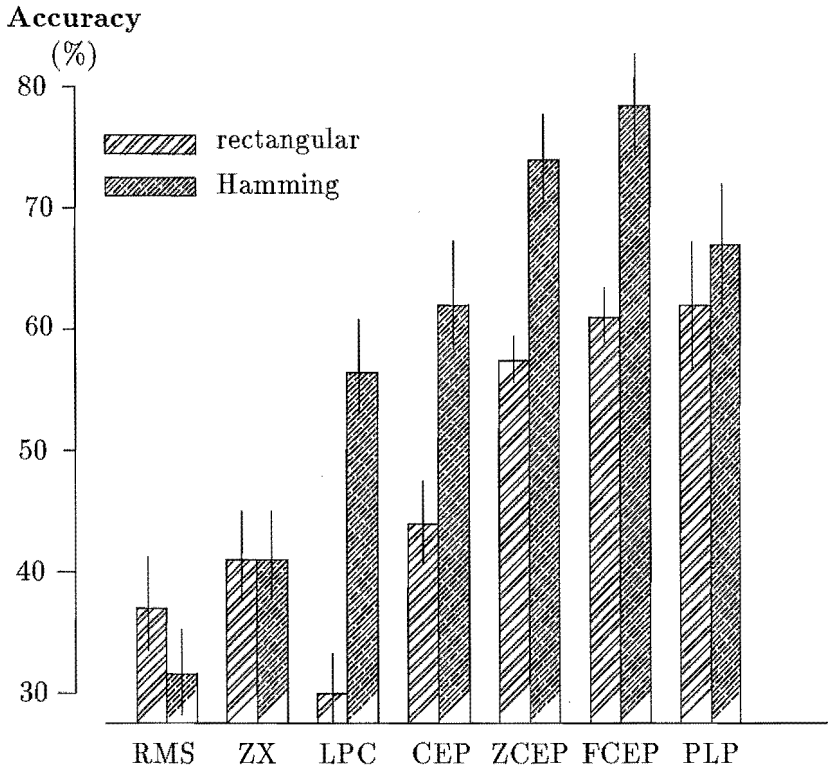


Figure 8.9. Average recognition accuracy and standard deviation (shown as a vertical line at the top of each bar) of the recognition accuracy for each feature when using Hamming and rectangular windowing. Results averaged over all speakers. Results are obtained for a UE21 DTW method with constant weightings. Tests performed on New Zealand speakers in speaker-dependent mode with 10 reference templates per word. A Euclidean distance is used.

windowing with a Hamming window gives lower LPC prediction error for frame sizes of two to three pitch periods. As the frame size increases the effect of window type becomes less important.

Examining Fig. 8.11 shows that the Hamming window distorts the LPCs greater than the rectangular window for small window sizes. However it is important to note that the recognition accuracy when using the Hamming window is still significantly higher than the rectangular window's recognition accuracy at all frame sizes.

Accuracy differences between Blackman, Hanning and Hamming window were also tested. Although not given here, the results showed no significant difference in recognition accuracy between these window types.

8.2.1.3 Pre-emphasis

To accurately model the spectral properties of the vocal tract, that is without the effects of glottal waveform and lip radiation characteristics, then it is important to pre-emphasise the speech before analysis (Markel and A.H. Gray, 1976). Pre-emphasis has the effect of boosting the lower amplitude high frequency component of the speech hence flattening the spectrum. This spectral flattening is more important for voiced speech, which has an approximate 6 dB per octave decrease with respect to frequency. Unvoiced sounds, which have a much flatter spectrum do not need this pre-emphasis. In fact it may grossly exaggerate the high frequency present if pre-emphasis is applied to unvoiced sounds. For this reason some researchers have proposed adaptive pre-emphasis, pre-

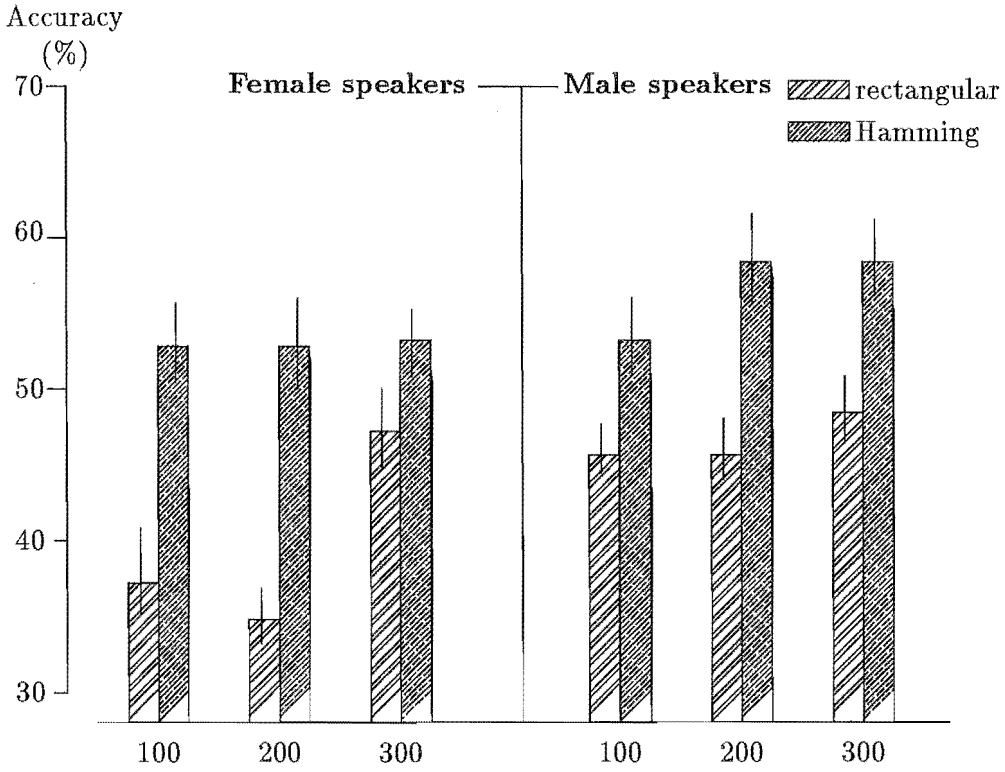


Figure 8.10. Average recognition accuracies and standard deviations (shown as a vertical line at the top of each bar) of the recognition accuracies calculated when using Hamming and rectangular windowing, for three different window sizes. Results for both male and female speakers are average over all the features.

emphasis dependent on the speech characteristics (Gray, Jr. and Markel, 1974). For practical word recognition systems the operation of an adaptive method is far too time consuming. Rather it may be more practical to see for which feature pre-emphasis is necessary and for which feature it is not necessary. Testing of this can be done by considering the recognition accuracies for both pre-emphasised and non emphasised data. The pre-emphasis technique is implemented as a simple one-zero filter of the form,

$$1 - ae^{-j2\pi fT}, \tag{8.2}$$

where f is frequency and T is the time between samples. The constant a is a scale factor determining the roll-off of the filter. For most systems the scale factor is set between 0.9 to 1.0, and in this case 0.95 is chosen. The operation of the filter on the speech is equivalent to simply differencing the sampled data, $s(n)$, in the time domain so that the pre-emphasised data, $s_p(n)$ can be found by,

$$s_p(n) = s(n) - 0.95[s(n-1)]. \tag{8.3}$$

Fig. 8.12 shows results for individual speakers, while Fig. 8.13 shows the recognition accuracies obtained for each feature with and without pre-emphasis. Examining the speaker's data from Fig. 8.12, male speakers showed a slightly greater improvement when pre-emphasis was used than female speakers with an average improvement of

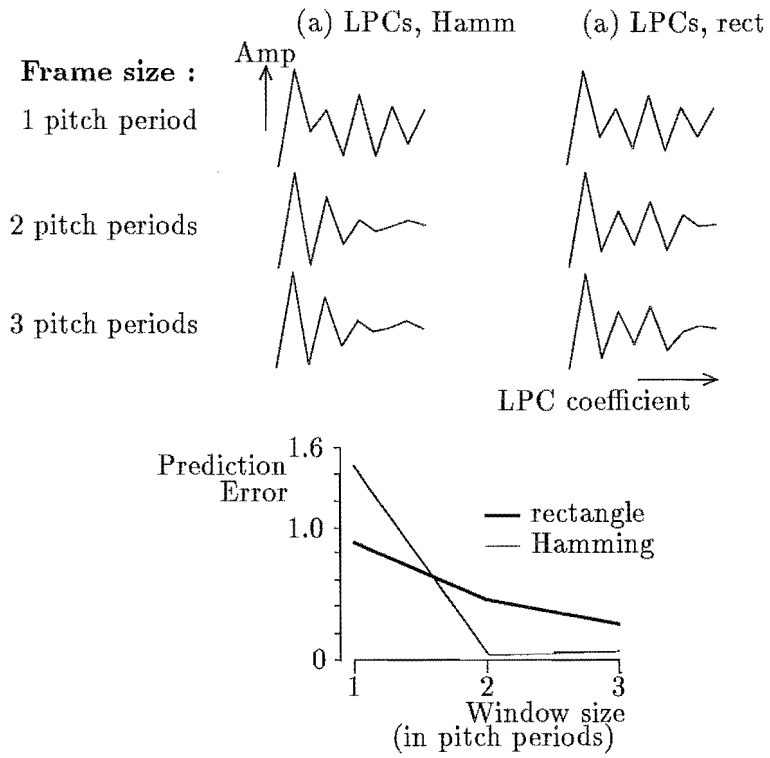


Figure 8.11. Comparison of LPC features for differing frame sizes. Window placement is pitch synchronous and at the beginning of pitch period.

10.4% for males and 10.2% for females. Although there is no significant difference between the male and female results both, male and female results, show significant improvement over no pre-emphasis. As shown in Fig. 8.13, pre-emphasis was much more important for ZX than for other features. For RMS pre-emphasis actually reduced performance. RMS accuracy decreased on average 5% for male speakers and 2% for female speakers. Several of the other features showed no significant change. On average the recognition accuracies for ZX improved by 49.6% for male speakers and 37.3% for female speakers. LPC accuracy increased only 6.8% for males and a 5.2% for female speakers.

The decrease in accuracy for RMS after pre-emphasis may be dependent on the vocabulary chosen. Because RMS models the amplitude variation of the speech waveform, increasing the amplitude, by pre-emphasis, of the higher frequency regions, causes greater confusion. This is because the (usually) lower amplitude unvoiced parts which contain the high frequencies are now boosted in amplitude increasing the confusion between words which were distinguishable by their differences in voiced and unvoiced amplitudes. With a digit vocabulary of many single syllable words this effect increases the confusion particularly with such words as ZERO, ONE, TWO, THREE, FOUR, and FIVE. Fig. 8.14 shows examples of the amplitudes of these words before and after pre-emphasis illustrating how the similarities increase. As this figure illustrates, similarities between the word sets (ZERO, ONE), (TWO, THREE, FOUR) and (ONE, FIVE) increase.

No relationship was observed between pre-emphasis and window type or frame size.

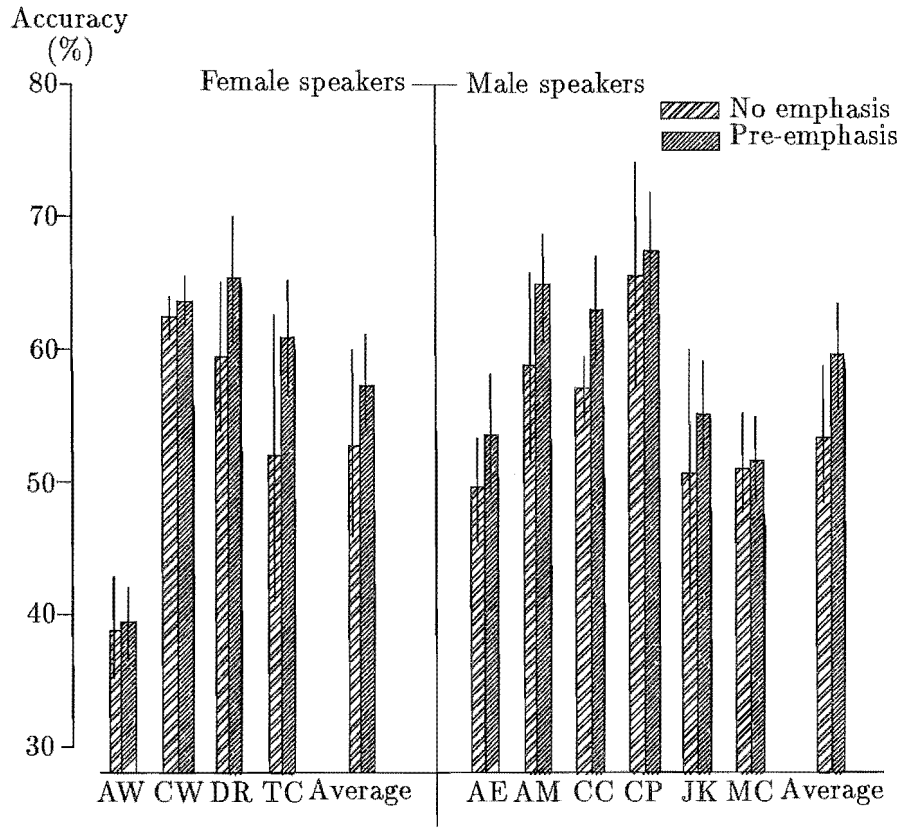


Figure 8.12. Average recognition accuracies and standard deviation (shown as a vertical line at the top of each bar) of the recognition accuracies averaged over all features for pre-emphasised and non emphasised speech. Plots show accuracy for each individual speaker and the average for male and female speakers. Recognition tests used a UE21 DTW method with constant weightings. Tests were with New Zealand speakers in speaker-dependent mode. 10 reference templates were stored per word and a Euclidean distance measure was used.

8.2.1.4 Overlapping data frames

If window frames are not overlapped information losses can occur when calculating features from the speech. This is particularly so if the window used on the speech frame shapes the speech by slowly reducing to zero at the edges of the frame. Usually the information near the edges of the frames can be retained by overlapping the adjacent frames. In order to ascertain how important it is to retain this information, recognition performances for frame overlaps of 12%, 25% and 40% of the frame size were tested. No significant improvement in recognition accuracy was observed from these tests for window frames, for either a Hamming or rectangular window, overlapped by 12% and 25%. In fact, the recognition accuracies decreased when a frame overlap of 40% was used.

Individual feature analysis for data frame overlap showed no significant change in accuracy for any features except ZX, which showed slight accuracy improvement.

These results showed that not only is the overlapping of frames unnecessary, but that it may also decrease the accuracy of recognition if overlap is too large.

The reason why accuracy may not have significantly increased is not known, however it may point to the affect that accurately representing all variation in the speech signal may not be useful for recognition. Accurate representation may introduce variations

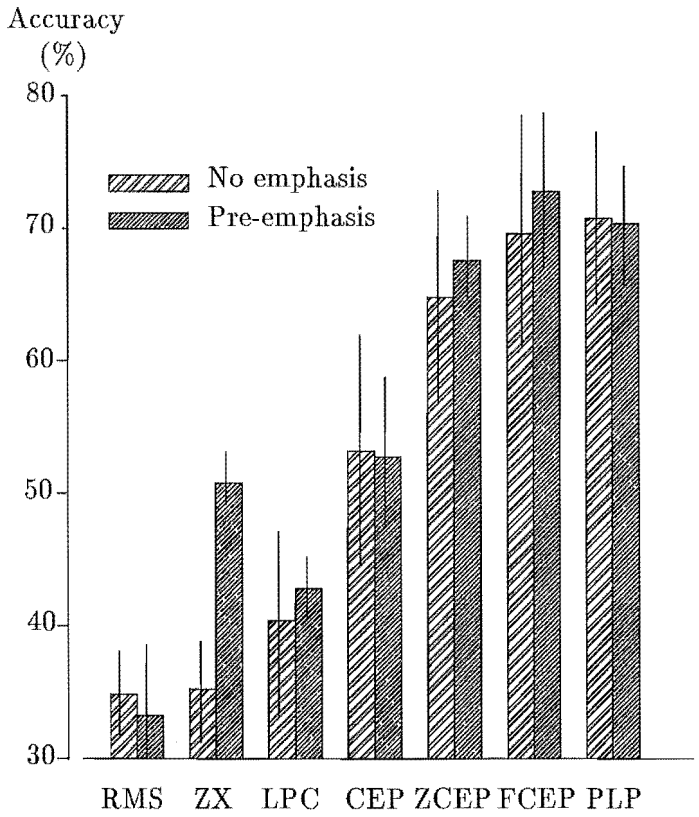


Figure 8.13. Average recognition accuracies and standard deviation (shown as a vertical line at the top of each bar) of the recognition accuracies for pre-emphasised and non emphasised speech averaged over all speakers (male and female) for each feature tested. Recognition tests used a UE21 DTW method with constant weightings. Tests were with New Zealand speakers in speaker-dependent mode. 10 reference templates were stored per word and a Euclidean distance measure was used.

which are not replicated from one reference word to another. In such cases recognition may be more successful when only the largest variation are modelled.

8.2.1.5 Number of Reference Templates

The number of reference templates has a major impact on the computation requirements and also effects the recognition performance. Thus, it is important to determine the optimum method of reference template selection to obtain the least number of reference templates for the task.

Two methods of reference template training were tested. These methods, known as the *casual training method* and the *statistical clustering method* are outlined in §7.5. The outcome of the recognition tests employing each of these methods is described in the following paragraphs.

To determine the effect that the number of reference templates has on the recognition accuracy, tests with 2, 6 and 10 templates were examined with the casual training method. Both speaker-dependent and speaker-independent tests were carried out. Both tests involved male and female speakers from the New Zealand database and speaker-independent tests from the American databases. A UE21 DTW recognition method with non-constant weightings, as discussed in §5.1.2.4, was employed. The method of recognition decision in all cases was the nearest neighbour approach. Reference and

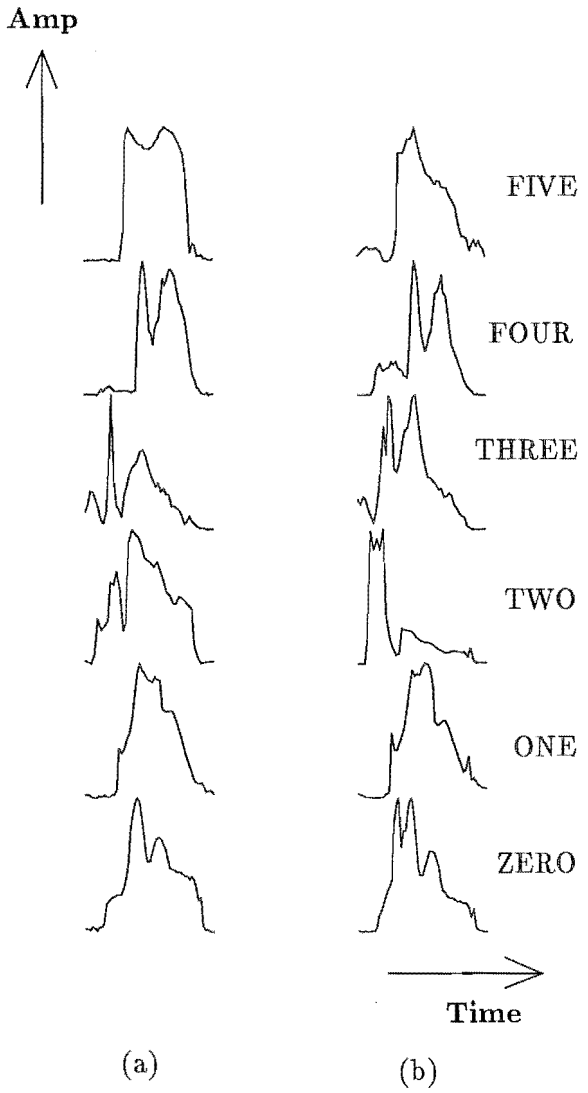


Figure 8.14. RMS plots of (a) Non emphasised and (b) pre-emphasised speech showing similarities between words (ZERO, ONE), and (TWO, THREE and FOUR), and (ONE, FIVE).

test data was jackknifed (§8.1.5.1) and 15 independent test sets from the same data was used, that is the data was rotated only 5 times. This meant that when testing with ten reference words 15 sets of 10 tests per word were undertaken, giving 1500 tests altogether. For 6 reference templates 15 independent test sets were run with 14 tests per word, giving 2100 recognition tests altogether, and for 2 reference templates 2700 recognition tests were run. This allowed the estimation of means and variations of the recognition results, as shown in the following tables. Tables 8.1, 8.2, and 8.3 also show the variation in the accuracy (the minimum and maximum accuracies) caused by changing reference templates from 2 to 6 to 10 templates for each of the words, while the average accuracies from these tests are given in Table 8.4 and Table 8.5. Note that for some of the speakers, maximum recognition accuracies shown in Tables 8.1, 8.2, and 8.3, and average accuracies shown in Tables 8.4 and 8.5, decreased when the number of templates increased from two to ten. For most features tested, accuracy increased significantly as template numbers increased. In such cases greater accuracy is expected because word variations are being represented with greater precision.

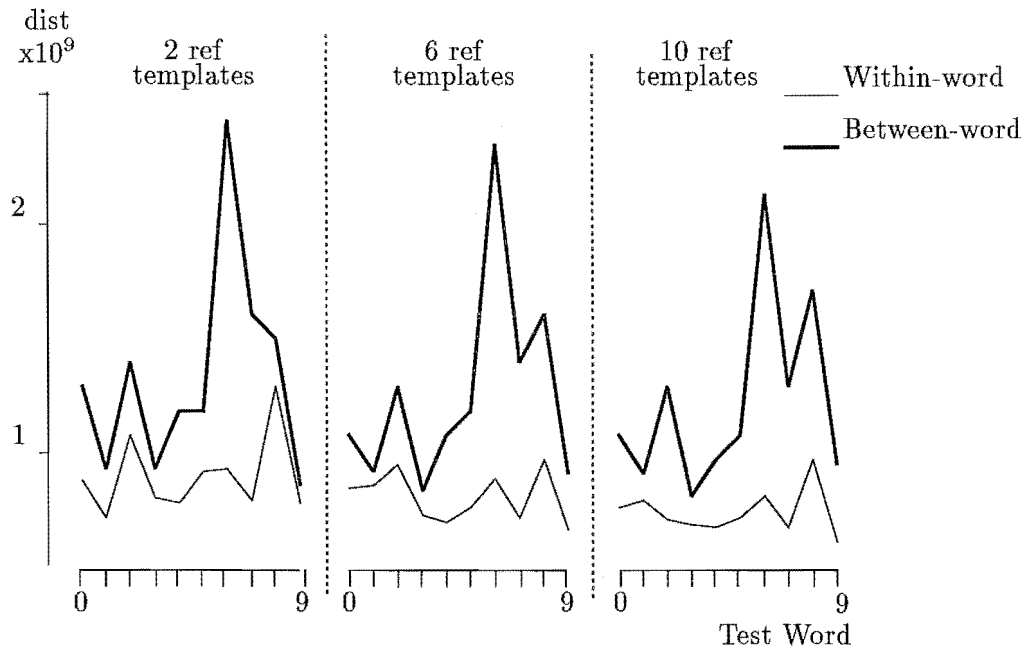
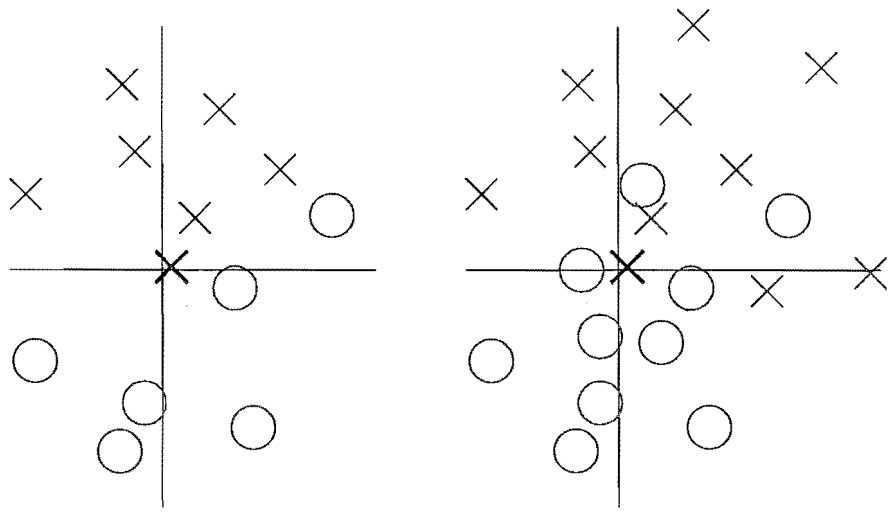


Figure 8.15. The effect on average between-word and average within-word distances for cepstral coefficients when the number of reference templates is increased. Reference templates increase from (a) 2 representations, (b) 6 representations, (c) 10 representations per word. Plots show average within-word and between word distances with averages calculated from the entire NZ database (20 speakers speaking 20 representations each of the ten words).

Since the ability of any recognition scheme can be estimated by its ability to separate the right word from the wrong words, it is interesting to examine the within-word and between-word distances. Fig. 8.15 illustrates the difference in the average between-word distances and the average within-word distances (refer §6.2) when template numbers are increased from 2 to 6 to 10 representations for cepstral coefficients. The averages were calculated from all the word tests for all the speakers tested. As the number of representations increase the average between-word and within-word distances decrease, however the average within-word distance decrease is greater than the average between-word distance decrease thereby increasing the separation between the wrong and the right word representations thereby producing higher recognition accuracies.

An accuracy increase when more reference templates are used is not necessarily the case for all feature types. There was a decrease in mean accuracy for the speaker dependent NZ male PLP accuracy between 6 to 10 templates (refer table 8.5). There was also a decreases in the maximum accuracy for a number of features In most cases the decrease occurred in speaker-independent mode and generally decreased when the template number was increased from 6 to 10 using the casual training method. Reductions actually occurred with RMS (NZ male speaker dependent), ZX (NZ female speaker independent, NZ male speaker independent, 2-6 templates), LPC (NZ female speaker independent), ZCEP (NZ female speaker independent), FCEP (NZ male speaker independent), and PLP (US female speaker independent and NZ male speaker independent). These decreases in average and maximum accuracy, caused by a greater numbers of templates, may be due to larger numbers of outliers, or by multiple representation of the different words overlapping more causing greater confusion. These affects are illustrated in Fig. 8.16.



(a) 6 templates per word (b) 10 templates per word

Figure 8.16. A possible outcome of increasing reference template numbers from 6 representations per word to 10 representations per word. Two words are represented by x and o in two-dimensional space. The test word **x** is correctly recognised (either when the reference representations are clustered or not) as word type x when 6 representations are used, however it is incorrectly recognised as o when 10 representations are used.

Large fluctuations of accuracy were observed when sets of reference templates changed by using a jackknife method on the data (refer §8.1.5.1). Fluctuations of accuracies up to 30% for some features occurred, and can be seen from the variations of accuracy tabulated in Tables 8.1, 8.2, and 8.3. These tables show the effect of template selection on recognition accuracy when reference templates are chosen in a random manner using the casual training method. With this method of training, unless an exhaustive pretest on the reference templates is run prior to template selection, an optimal set of templates cannot be chosen. Running such a pre-test would be equivalent to running multiple recognition tests to find the best template selection.

The problem of reference template selection causing wide accuracy fluctuations was first noted by Russell, Deacon and Moore in 1984. They showed the same as what has been shown in this thesis, that testing with only one set of reference templates can not give an accurate estimation of a recognition system's (where a recognition system involves all pre and post processing of a speech sample including sampling, training, recognition processes, etc) true capability. A systems true capability is its predictable capability, where this capability must, therefore, consist of and average (expected) recognition accuracy and a variance. Since a true capability cannot be found in this manner (using only a limited sample set) it is also erroneous to compare different systems which are tested on different, limited, reference vocabularies if exhaustive testing has not been done. Obviously, to accurately characterise a recognition system's ability it is imperative to undertake tests which give statistical variations.

Because it is difficult to obtain accurate, repeatable, results with randomly selected reference templates, accuracy may be increased, and variance decreased when templates are selected by some statistical process such as clustering. In the following section both random selection and cluster are examined.

The testing for the two methods of template selection in this thesis was done by pro-

Speaker-independent, NZ male speakers

			Templates/word		
Parameter	Distance	Training	2	6	10
RMS	Euclid	*CT*	21.7-14.4	23.1-18.6	24.0-20.0
		SC	18.8-10.0	18.8-8.8	
ZX	Euclid	*CT*	50.0-30.0	42.9-36.0	44.0-35.0
		SC	33.8-10.0	42.5-11.3	
LPC	Euclid	*CT*	54.4-30.0	64.2-48.3	67.8-54.4
CEP	Euclid	*CT*	63.3-34.4	75.0-66.2	79.0-63.5
		SC	86.3-67.5	86.3-68.0	
ZCEP	Euclid	*CT*	64.0-33.3	70.0-62.3	73.0-60.0
		SC	80.0-62.5	80.0-61.3	
FCEP	Euclid	*CT*	57.2-32.7	69.8-55.7	67.0-51.3
		SC	71.3-55.0	71.3-56.3	
PLP	Euclid	*CT*	30.9-29.5	44.4-32.2	44.1-34.2
CEP	WEuclid	*CT*	69.4-33.9	73.8-65.7	78.0-57.5
CEP	projection	*CT*	72.2-55.5	77.1-65.5	78.0-63.0

Speaker-independent, NZ female speakers

			Templates/word		
Parameter	Distance		2	6	10
RMS	Euclid	*CT*	26.3-17.5	32.4-20.0	40.0-35.0
		SC	14.4-5.6	21.1-12.2	
ZX	Euclid	*CT*	28.8-17.8	42.6-24.1	30.0-25.0
		SC	16.7-10.0	18.8-14.4	
LPC	Euclid	*CT*	83.1-45.0	87.5-65.0	85.0-78.3
CEP	Euclid	*CT*	83.8-58.5	87.5-71.7	90.0-83.8
		SC	74.4-67.8	91.1-88.9	
ZCEP	Euclid	*CT*	81.3-52.5	89.2-72.5	88.8-75.0
		SC	72.2-63.3	78.8-65.6	
FCEP	Euclid	*CT*	65.0-45.0	70.0-58.3	73.8-65.0
		SC	63.3-56.7	81.1-72.2	
PLP	Euclid	*CT*	18.8-18.3	40.0-34.3	48.0-33.0
CEP	WEuclid	*CT*	88.1-59.4	93.3-77.5	93.3-83.8
CEP	projection	*CT*	78.1-55.6	85.0-73.3	90.0-75.6

Table 8.1. Recognition accuracy variations with differing reference template selection, either using the casual training method, *CT*, or the statistical clustering method, *SC*. Results show maximum and minimum speaker-independent recognition accuracies. The variation of accuracy is calculated by changing the reference and test templates using the jackknife method. Various recognition distance methods are also trialed. The distance methods include Euclidean, weighted Euclidean (WEuclid), and cepstral projection (projection). Results for New Zealand male and female speakers.

Speaker-independent, US male speakers

			Templates/word		
Parameter	Distance		2	6	10
RMS	Euclid	*CT*	30.0-14.0	33.1-23.1	38.3-29.2
		SC	28.8-10.0	31.3-13.8	
ZX	Euclid	*CT*	43.8-27.5	45.8-31.7	53.3-41.1
		SC	30.0-7.5	38.8-10.0	
LPC	Euclid	*CT*	66.0-37.5	75.6-63.1	77.5-63.3
CEP	Euclid	*CT*	73.5-45.5	78.1-69.4	81.7-63.3
		SC	78.8-59.7	83.8-59.7	
ZCEP	Euclid	*CT*	70.0-51.5	82.5-65.0	85.8-74.2
		SC	80.0-63.5	86.3-64.9	
FCEP	Euclid	*CT*	64.5-41.5	72.5-60.6	75.8-62.5
		SC	78.8-61.3	77.5-62.5	
PLP	Euclid	*CT*	36.8-31.8	57.5-52.5	60.0-52.5
CEP	WEuclid	*CT*	78.0-48.0	85.0-72.5	91.6-76.0
CEP	projection	*CT*	76.5-50.0	86.3-70.6	90.8-75.8

Speaker-independent, US female speakers

			Templates/word		
Parameter	Distance		2	6	10
RMS	Euclid	*CT*	31.7-14.4	39.2-27.9	42.0-30.0
		SC	22.5-11.3	25.0-12.5	
ZX	Euclid	*CT*	52.2-36.7	55.0-43.6	68.0-45.0
		SC	32.5-12.5	38.8-20.0	
LPC	Euclid	*CT*	68.3-48.9	71.4-61.4	79.0-63.0
CEP	Euclid	*CT*	70.0-55.5	71.4-60.7	83.0-62.0
		SC	81.3-55.0	81.3-62.5	
ZCEP	Euclid	*CT*	70.6-58.3	79.3-67.9	86.0-63.0
		SC	87.5-58.8	87.5-71.3	
FCEP	Euclid	*CT*	69.4-56.1	75.0-60.0	80.0-50.0
		SC	80.0-63.8	80.0-66.3	
PLP	Euclid	*CT*	45.0-39.3	62.5-56.7	61.3-53.7
CEP	WEuclid	*CT*	63.1-26.3	72.5-59.1	82.2-66.7
CEP	projection	*CT*	73.3-60.0	75.7-67.1	81.0-65.0

Table 8.2. Recognition accuracy variations with differing reference template selection, either using the casual training method, *CT*, or the statistical clustering method, *SC*. Results show maximum and minimum speaker-independent recognition accuracies. The variation of accuracy is calculated by changing the reference and test templates using the jackknife method. Various recognition distance methods are also trialed. The distance methods include Euclidean, weighted Euclidean (WEuclid), and cepstral projection (projection). Results for American male and female speakers.

Speaker-dependent, NZ male speakers

			Templates/word		
Parameter	Distance		2	6	10
RMS	Euclid	*CT*	54.3-36.4	66.0-45.0	70.0-56.7
		SC	73.3-10.0	68.3-20.0	
ZX	Euclid	*CT*	62.1-51.4	78.0-56.0	88.0-66.7
		SC	20.0-13.0	50.0-30.0	
LPC	Euclid	*CT*	80.0-69.3	88.0-76.0	91.7-78.3
CEP	Euclid	*CT*	91.4-76.4	93.0-87.0	98.3-91.7
		SC	93.0-75.0	98.3-88.3	
ZCEP	Euclid	*CT*	82.1-69.3	89.0-79.0	95.0-80.0
		SC	88.3-71.7	88.3-78.3	
FCEP	Euclid	*CT*	77.9-60.7	88.0-73.0	95.0-81.7
		SC	78.3-75.3	86.7-76.7	
PLP	Euclid	*CT*	76.7-61.4	85.0-58.5	88.0-52.8
CEP	WEuclid	*CT*	91.4-77.1	93.0-82.0	95.0-88.3
CEP	projection	*CT*	88.6-75.0	91.0-85.0	98.3-90.0

Speaker-dependent, NZ female speakers

			Templates/word		
Parameter	Distance		2	6	10
RMS	Euclid	*CT*	52.9-35.0	65.0-52.0	70.0-55.0
		SC	26.7-10.0	51.6-26.7	
ZX	Euclid	*CT*	62.1-52.8	76.0-49.0	83.3-35.0
		SC	15.0-10.0	48.3-11.7	
LPC	Euclid	*CT*	89.7-77.9	92.0-87.0	98.3-83.3
CEP	Euclid	*CT*	92.1-85.0	97.0-93.0	100.-95.0
		SC	95.0-73.3	98.3-95.0	
ZCEP	Euclid	*CT*	91.4-82.1	97.0-91.0	100.-90.0
		SC	88.3-63.3	98.3-93.3	
FCEP	Euclid	*CT*	83.6-72.8	95.0-83.0	100.-78.3
		SC	83.3-66.7	96.7-73.3	
PLP	Euclid	*CT*	44.0-40.0	60.0-55.5	81.4-71.4
CEP	WEuclid	*CT*	95.0-85.7	100.-95.0	100.-95.0
CEP	projection	*CT*	95.7-87.1	98.0-93.0	100.-93.3

Table 8.3. Recognition accuracy variations with differing reference template selection, either using the casual training method, *CT*, or the statistical clustering method, *SC*. Results show maximum and minimum speaker-dependent recognition accuracies. The variation of accuracy is calculated by changing the reference and test templates using the jackknife method. Various recognition distance methods are also trialed. The distance methods include Euclidean, weighted Euclidean (WEuclid), and cepstral projection (projection). Results for New Zealand male and female speakers.

Speaker-independent, male NZ			templates/word	
Parameter	Distance	2	6	10
RMS	Euclid	18.4	20.7	22.1
ZX	Euclid	40.2	40.1	40.3
LPC	Euclid	47.9	57.9	63.3
CEP	Euclid	54.3	68.2	71.9
ZCEP	Euclid	53.5	65.5	67.7
FCEP	Euclid	48.2	59.9	62.0
PLP	Euclid	30.0	38.9	40.2
CEP	WEuclid	57.7	69.0	70.0
CEP	projection	62.2	69.2	75.0
Speaker-independent, female NZ			templates/word	
Parameter	Distance	2	6	10
RMS	Euclid	21.1	26.6	37.1
ZX	Euclid	25.3	31.2	26.7
LPC	Euclid	62.5	74.7	80.7
CEP	Euclid	69.5	79.6	86.3
ZCEP	Euclid	66.1	78.9	82.3
FCEP	Euclid	56.0	64.4	70.2
PLP	Euclid	18.5	37.6	39.7
CEP	WEuclid	74.3	85.4	91.8
CEP	projection	67.9	79.4	81.5
Speaker-independent, male US			templates/word	
Parameter	Distance	2	6	10
RMS	Euclid	23.7	28.9	33.0
ZX	Euclid	35.3	37.4	47.2
LPC	Euclid	53.5	67.1	70.5
CEP	Euclid	60.8	72.6	74.3
ZCEP	Euclid	60.2	73.6	79.6
FCEP	Euclid	53.5	66.1	70.8
PLP	Euclid	34.3	55.0	56.3
CEP	WEuclid	62.5	78.5	82.6
CEP	projection	61.4	78.7	82.6
Speaker-independent, female US			templates/word	
Parameter	Distance	2	6	10
RMS	Euclid	27.4	33.9	34.1
ZX	Euclid	45.3	49.9	52.2
LPC	Euclid	57.4	66.3	71.1
CEP	Euclid	62.3	67.9	72.4
ZCEP	Euclid	64.8	71.6	75.1
FCEP	Euclid	61.7	65.3	66.1
PLP	Euclid	45.0	58.7	57.4
CEP	WEuclid	50.3	67.5	71.7
CEP	projection	66.3	70.2	75.8

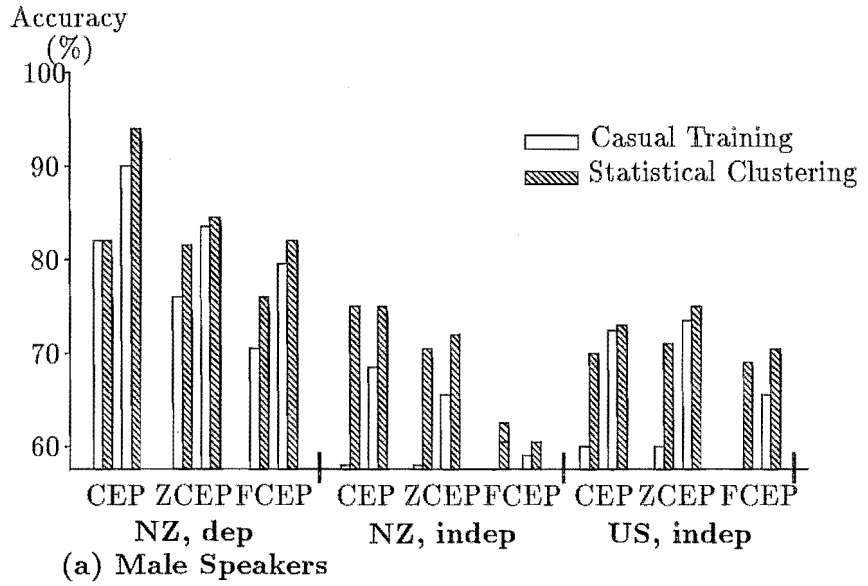
Table 8.4. Average recognition accuracies for speaker-independent male and female New Zealand and American accented speakers with differing numbers of reference templates (either 2, 6, or 10) using the casual training method. The averages are calculated by changing the reference and test templates using the jackknife method. Various recognition distance methods are also trialed. The distance methods include Euclidean, weighted Euclidean (WEuclid), and cepstral projection (projection).

Speaker-dependent, male NZ			templates/word	
Parameter	Distance	2	6	10
RMS	Euclid	43.9	55.3	61.0
ZX	Euclid	56.6	67.0	75.3
LPC	Euclid	74.1	80.7	86.3
CEP	Euclid	82.0	90.7	94.8
ZCEP	Euclid	76.6	83.4	88.5
FCEP	Euclid	70.6	79.5	86.3
PLP	Euclid	69.0	70.8	66.9
CEP	WEuclid	83.0	88.4	91.4
CEP	projection	82.5	89.0	95.0
Speaker-dependent, female NZ			templates/word	
Parameter	Distance	2	6	10
RMS	Euclid	44.1	57.6	64.0
ZX	Euclid	57.4	62.3	71.8
LPC	Euclid	84.4	89.1	92.2
CEP	Euclid	88.8	94.7	97.5
ZCEP	Euclid	86.4	93.5	96.2
FCEP	Euclid	79.8	88.3	93.3
PLP	Euclid	41.5	58.0	78.5
CEP	WEuclid	91.2	97.1	99.2
CEP	projection	88.6	96.2	97.7

Table 8.5. Average recognition accuracies for speaker-dependent male and female New Zealand accented speakers with differing numbers of reference templates (either 2, 6, or 10) using the casual training method. The averages are calculated by changing the reference and test templates using the jackknife method. Various recognition distance methods are also trialed. The distance methods include Euclidean, weighted Euclidean (WEuclid), and cepstral projection (projection).

ducing a reduced set of reference templates from an original set of reference templates. The original reference templates for both methods were chosen from the same set of ten templates. Either clustering or random selection is used to reduced the number of templates to 6 or 2. Many trials were performed for each test by changing the original set of ten templates using a jackknife method (refer §8.1.5.1). From the many trials the mean accuracies are calculated and given in Table 8.6 and Table 8.7 for the clustering method. Table 8.6 and 8.7 can be compared with the mean accuracy results given in Tables 8.4 and 8.5 for the random selection method.

Recognition accuracies are generally higher when words are clustered to fewer templates. Fig. 8.17 shows average recognition accuracies when reference templates are clustered and when random selection is used. From this figure and tables 8.1,8.2, and 8.3 it can be seen that (except for RMS and ZX) accuracies are up to 300% higher when clustering is used and that the highest increases were for CEP features. Thus for CEP and transitional CEP it would be better to cluster the speech data while for RMS and ZX, random selection is best.



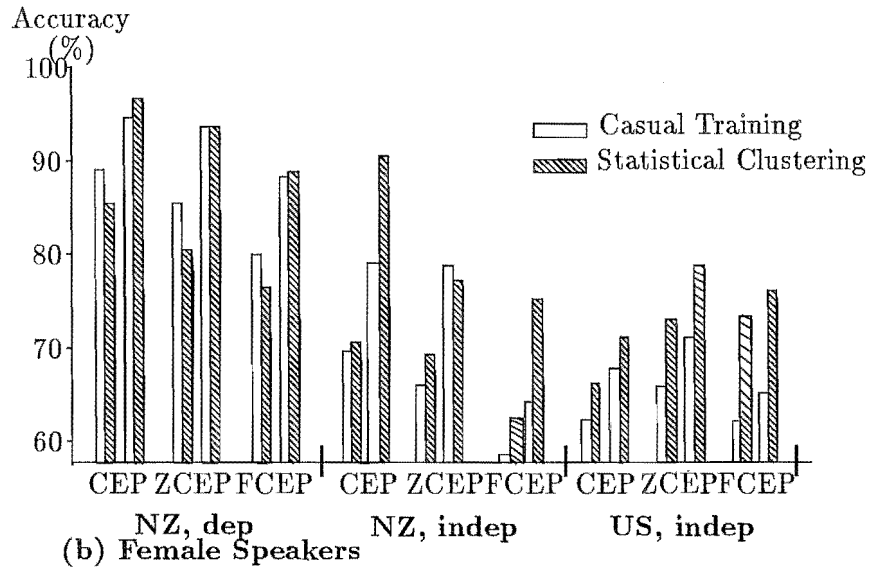


Figure 8.17. Recognition accuracies using two different methods of template selection; the casual training method and the statistical clustering method. Recognition accuracies for three different features (cepstral (CEP), and transitional cepstral (ZCEP and FCEP)) are shown for both (a) male and (b) female speakers in speaker-dependent and speaker-independent modes for US and NZ accents. Results for differing numbers of reference templates, 2 or 6 (first and second members of each feature group), is also shown.

8.3 TESTING OF RECOGNITION FEATURES

One of the major problems of word recognition is knowing what is the best information to extract from the sampled speech to obtain highly accurate recognition. In this section many features, first discussed in Chapters 4 and 7, will be tested for their abilities for speech recognition.

The testing uses a UE21 DTW method with non-constant weightings (refer §7.6.2). Results are given with ten representations of each word used as reference templates, unless otherwise stated. Reference and test data are rotated using the jackknife method (§8.1.5.1). For the New Zealand speakers, both speaker-dependent and speaker-independent

Speaker-independent, male, NZ			
Parameter	Distance	2 templates/word	6 templates/word
RMS	Euclid	15.0	15.2
ZX	Euclid	18.8	21.1
CEP	Euclid	75.0	75.2
ZCEP	Euclid	70.9	72.9
FCEP	Euclid	62.2	61.5
Speaker-independent, female,NZ			
Parameter	Distance	2 templates/word	6 templates/word
RMS	Euclid	11.9	17.2
ZX	Euclid	12.7	20.2
CEP	Euclid	70.4	90.8
ZCEP	Euclid	69.4	77.4
FCEP	Euclid	62.4	75.6
Speaker-independent, male, US			
Parameter	Distance	2 templates/word	6 templates/word
RMS	Euclid	16.5	21.6
ZX	Euclid	18.4	21.3
CEP	Euclid	70.0	73.1
ZCEP	Euclid	71.5	75.0
FCEP	Euclid	68.9	70.4
Speaker-independent, female, US			
Parameter	Distance	2 templates/word	6 templates/word
RMS	Euclid	17.3	19.7
ZX	Euclid	20.0	26.4
CEP	Euclid	66.4	71.1
ZCEP	Euclid	73.1	79.4
FCEP	Euclid	73.3	76.1

Table 8.6. Average recognition accuracies for speaker-independent male and female New Zealand and American accented speakers with differing numbers of reference templates (either 2 or 6) using the statistical training method. The averages are calculated by changing the reference and test templates using the jackknife method.

tests are trialed with reference templates chosen using the casual training method (refer §7.5). For the American speakers, only speaker-independent tests are performed with the reference templates chosen using the casual training method.

8.3.1 Recognition Features used Individually

The features tested were chosen to cover a range of methods often discussed in the literature. These methods are both temporal, such as zero-crossing rate (ZX) and energy (RMS), and frequency based, such as predictor coefficients (LPC), cepstral coefficients (CEP) and perceptual coefficients (PLP). Dynamic cepstrals (ZCEP,FCEP) were also tested to discover if transitional information is advantageous for speech recognition. A discussion of how these features are derived is given in Chapters 4 and 7.

The minimum and maximum accuracies of each feature for the speaker-dependent tests on the New Zealand speakers are given in Table 8.3, with the mean accuracies

Speaker-dependent, male, NZ			
Parameter	Distance	2 templates/word	6 templates/word
RMS	Euclid	29.2	41.4
ZX	Euclid	14.9	40.3
CEP	Euclid	82.0	94.0
ZCEP	Euclid	81.7	84.3
FCEP	Euclid	76.7	82.3
Speaker-dependent, female, NZ			
Parameter	Distance	2 templates/word	6 templates/word
RMS	Euclid	19.3	37.1
ZX	Euclid	12.4	31.1
CEP	Euclid	86.4	97.2
ZCEP	Euclid	80.5	93.6
FCEP	Euclid	76.9	88.7

Table 8.7. Average recognition accuracies for speaker-dependent male and female New Zealand accented speakers with differing numbers of reference templates (either 2 or 6) using the statistical training method. The averages are calculated by changing the reference and test templates using the jackknife method.

given in Table 8.5. Fig. 8.18 also shows the average accuracies (variance is shown on these plots, drawn as a vertical bar) plotted for the male and female New Zealand speaker separately and averaged for the speakers. Both the cepstral coefficients and the transitional coefficients derived from the cepstral coefficients give the highest accuracy for these speaker-dependent tests. The transitional coefficient's accuracy is slightly higher for the female speakers than the cepstral coefficient's accuracy and vice-versa for the male speakers, although these differences are insignificant. For both the average male and average female speakers, LPC and PLP features give much lower accuracies than the maximum. The lowest accuracies occur with RMS and ZX.

Fig. 8.19 shows the recognition accuracy for each of the male speakers and female speakers individually. The individual recognition accuracies follow that of the average recognition accuracies. Cepstrals (CEP) and zeroth order transitional data (ZCEP) derived from cepstrals give the highest accuracies for each individual speaker. The LPC feature was able to equal the accuracy of CEP and ZCEP features for the speakers CW and CP. It can also be noted from this figure that the speakers with the highest accuracy for any particular feature do not have the highest accuracy for all features. Thus the relative performance of each speaker changes from feature to feature. The relative performance changes with feature type occur mainly with features that give low performances; those features that give high accuracies give high accuracies across all speakers.

Speaker-independent recognition performances are given in Table 8.4. These average results are also shown graphically in Fig. 8.20 with the variance of these results drawn as a vertical bar. The speaker-independent trials for both the USA and NZ speakers follow the same general trend as that of the speaker-dependent trials. Lowest accuracies were again from RMS and ZX, while CEP and ZCEP give the highest accuracies. Some accent dependent results are apparent from these results. The USA male and female ZCEP accuracies are significantly higher than the CEP accuracy while for the NZ speakers the CEP accuracies are higher than the ZCEP. Also, the USA speakers PLP accuracies are up to 100% greater than the NZ speakers' PLP accuracies suggesting

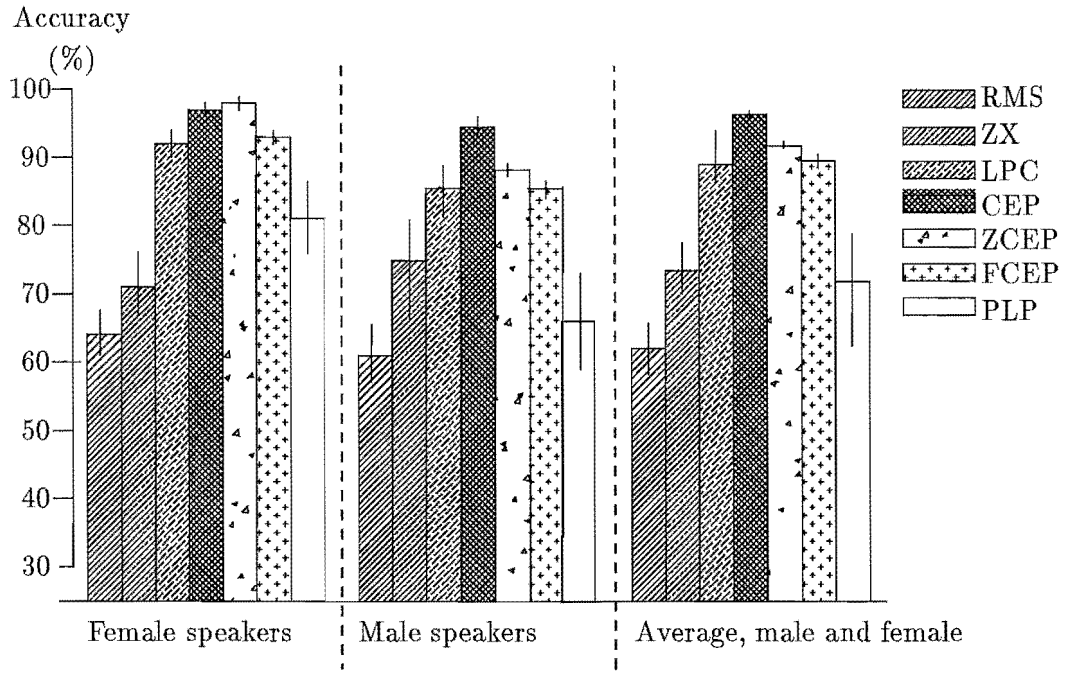


Figure 8.18. Speaker-dependent recognition accuracies for a range of features. Average recognition accuracies and standard deviations (shown as a vertical line at the top of each bar) of the recognition accuracies are given in the plot. Average results are calculated by averaging over all speaker-dependent New Zealand speakers' accuracies. The tests were carried out for both male and female speakers. Recognition trials used data which was initially pre-emphasised and Hamming windowed. For recognition 10 templates per word was used for training (using the casual training method) a UE21 DTW method with non-constant weightings.

accent dependencies with this feature. Overall the NZ accuracies were slightly higher than the American accuracies. This accuracy difference may be caused by a number of differences between the two databases. Some of the difference may include, for example, the way the databases were recorded, the different conditions which the databases were recorded under (including background noise, microphone, sampling etc), the difference in numbers of individual speakers, and the editing procedures which may have involved different manual endpoint detection.

8.3.2 Recognition Features Combined

The combination of individual features to improve recognition is a debatable issue. Areas which still require examination involve both the methods of combination and the choice of features to combine. In fact the combination of certain features, or particular methods of combination, may result in a degradation of accuracy. Many researchers have combined time based and frequency based measures such as zero crossings and energy (Lau and Chan, 1985), or LPC and energy (Rabiner *et al.*, 1984b) with only small or insignificant success. However, the combination of cepstral coefficients and transitional cepstral coefficients (Furui, 1986; Soong and Rosenberg, 1988; Furui, 1989; Hunt and Lefebvre, 1989) appears successful.

The following sections examine three distinct ways of combining features. The first method effectively combines the features during the DTW phase when calculating the local distance measure. This method allows the dynamic warping path to be altered

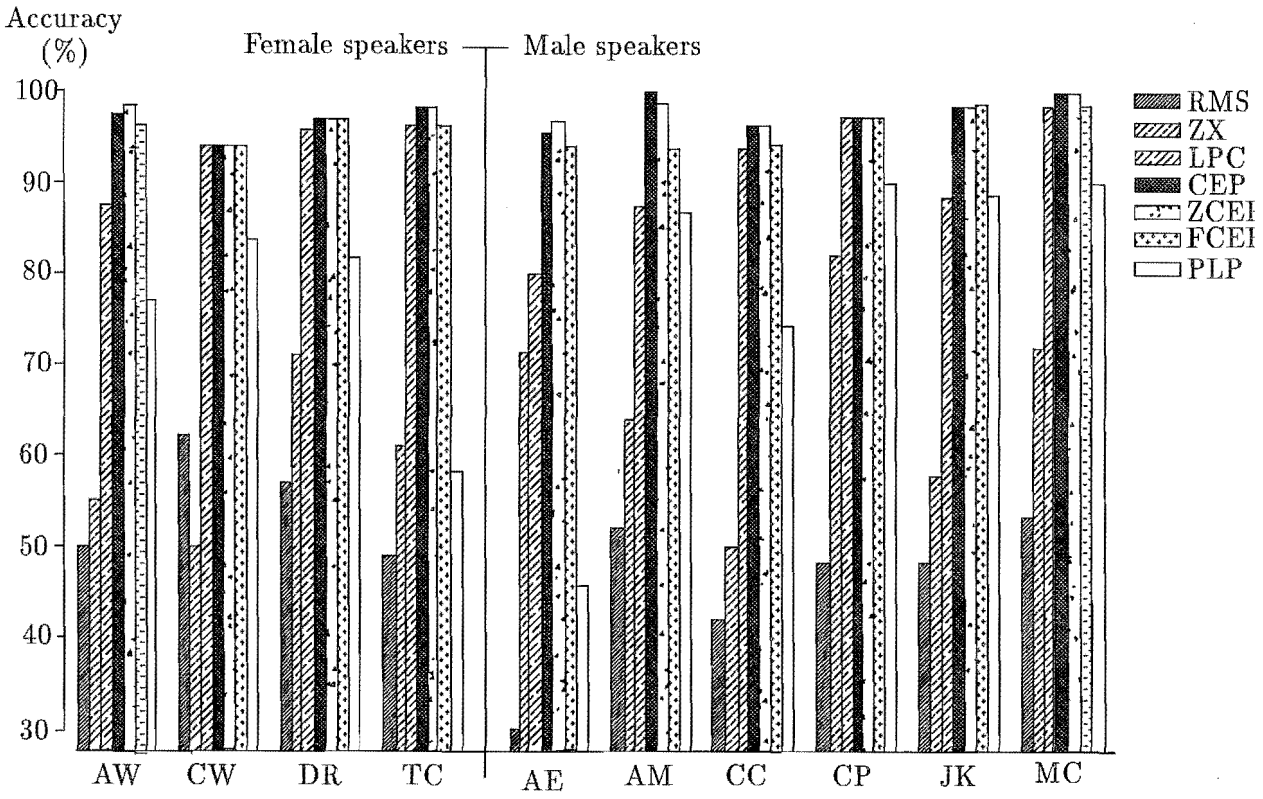


Figure 8.19. Individual speaker-dependent recognition accuracies for a range of features. Results are shown for the individual male and female New Zealand speakers. Recognition trials pre-emphasised and Hamming windowed the data. For recognition a UE21 DTW method is used with non-constant weightings.

with the inclusion of the extra parameter. The second method combines the sets of chosen word representations after each separate feature has been individually warped. A third method combines the sets of distance measures calculated from the DTW recognition after each feature has been separately warped.

8.3.2.1 Combining during DTW

Combining the features by combining each feature's distance which is calculated during the dynamic programming operation was suggested by Rabiner *et al*(1984b). For this method, therefore, one DTW warping path is calculated between a test and reference word no matter how many features are used. This method is particularly useful for combining two or more mutually exclusive features. When combining two mutually exclusive features during the dynamic time warping procedure, a warping path can be produced which is unlike either of the warping paths calculated separately by the two features. Two features that are often combined and that represent different information about the word, are LPCs and RMS. One difficulty, however, is that combining features with differing magnitude scales can allow one feature to swamp out another. In particular the combination of a multi-dimensional vector such as CEP or LPC with a one-dimensional vector such as RMS or ZX may also lead to magnitude problems. Either features or the feature's distances must be magnitude normalised before warping can take place. Another problem with the DTW method is that of the combination of the different features calculated distances at each grid point during the DTW calcula-

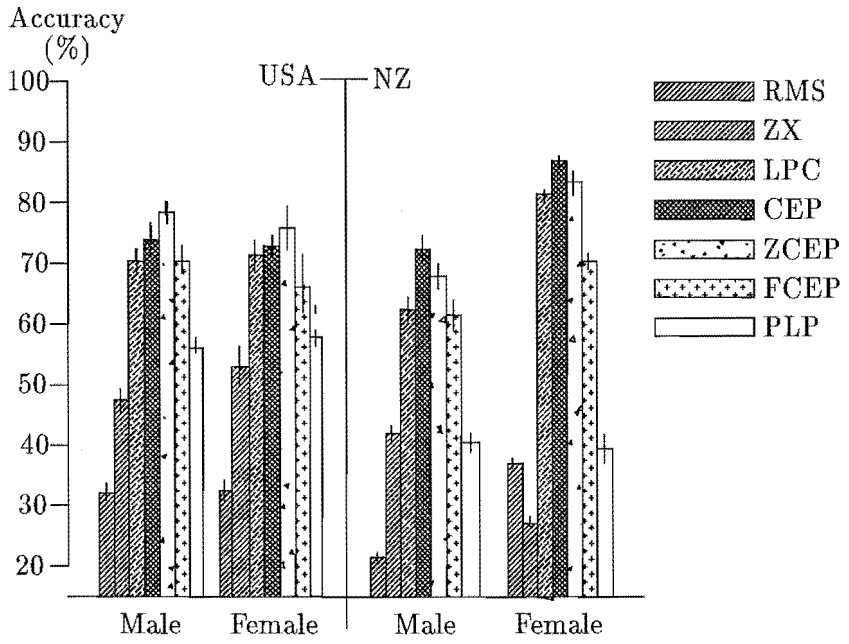


Figure 8.20. Average recognition accuracies and standard deviations (shown as a vertical line at the top of each bar) of the recognition accuracies for a range of features. Results for the speaker-independent tests with American and New Zealand male and female speakers using the casual training method with 10 templates per word. Pre-emphasis, Hamming windowing, and a UE21 DTW method with non-constant weightings were used.

tions. If each feature uses its own particular method of distance calculation to produce optimum performance and the combination of these features' distances can be difficult. One system discussed by Rabiner(1984) uses different distance measures for the two different parameters, namely the Itakura distance measure for the LPC parameters and a normalised log difference distance for the energy. The problem of combination is one of knowing the scale of each distance so to correctly apply any weightings to obtain an optimal combination. The scales may depend on variable of the speech such as intensity of speaking, or environmental factors such as background noise level. These difficulties are discussed in detail by Rabiner(1984) who claims a decrease of error up to 4.3%.

Using a DTW based method of feature combination an initial trial was undertaken for two speakers, in speaker-dependent mode. The two best recognition parameters, cepstral coefficients and transitional cepstral coefficients, were combined. For the female speaker a decrease in error of 24% was achieved (error rate dropped from 2.9% to 2.1%) but for the male speaker an increase of error rate from 2% to 8.3% occurred.

Based on all the initial tests the results did not show any significant accuracy improvement (using a t-test at 95% level) and so this method was abandoned.

8.3.2.2 Combining Word Choices

Another method of combining features is to examine the output sets of words selected by several independent recognition tests. The sets of words (or word list) from each feature consist of the top N choices, where N is known as the *depth*. The depth is chosen to ensure a 99% chance that the correct word resides in the set of chosen words. This probability of 99% is obtained by prior testing of the feature and examining the mean accuracy and standard deviations to the mean accuracy at various depths. N is

therefore dependent on the feature choice and speaker. For an accurate feature, such as CEP, N may be 5 while for an inaccurate feature N may be as large as 20. The word lists can then be unified in such a way as to emphasise members that occur more often, while de-emphasising those that occur least. Positional weighting on each word member, dependent on the word's overall depth in the list, allows those words that are higher and more often on the word list, to be weighted greater. This improves the recognition of a word which occurs often in the word list, but not necessarily the first choice. The total sum of all the words in all the lists to be combined is defined as M such that

$$M = \sum_{i=1}^{N_f} N_i \tag{8.4}$$

where N_i is the number of words in list i, and N_f is the number of features tested, then the individual weightings W_i for the words in a particular word list i from position $n_i = 1$ to $n_i = N_i$ are calculated as

$$W_{n_i} = (M/N_i).\frac{1}{n_i} \quad n_i = 1..N_i \tag{8.5}$$

and the selected word is that for which $\sum_{i=1}^{N_i} W_{n_i}$ is a minimum. A limitation of this method is that it requires that the accuracies of each individual feature be high so that the correct word choice exists often in the top positions. Another limitation is that the depth value, above which the algorithm takes its candidates, must be set prior to recognition. This depth value is speaker-dependent, and an accurate value for one speaker may not be accurate for another. The recogniser is therefore constrained to be a speaker-dependent one. To clarify this method the procedure can be expanded as follows;

1 For a set of features, such as RMS, CEP, *etc*, undertake DTW recognition for each feature individually and output word list. The word list i consists of N_i possible choices for the test word, where N_i is the depth of the list i and is dependent on the accuracy of the feature. The N_i choices are the top N_i candidates for the particular feature based on the calculated distances for that feature with every reference template.

2 Weight word choices in each word list based on position of word in word list given in equation (8.5).

3 Calculate a united word-list, where the word positions in the united list are found by summing up the positional weightings in the individual lists.

4 Final word choice is based on united list first word choice.

Initial testing of this combination method generally gave a lower combined recognition accuracy than the individual accuracies. The highest accuracy was obtained when combining cepstral and transitional cepstrals with individual accuracies of 91% (CEP), 96% (ZCEP), and 91% (FCEP) giving an average recognition result of 95% (with depth values of 6, 4 and 6 respectively).

This method continually gave lower accuracies then that of the best individual method. This is due to the combination of higher accuracy word lists with lower accuracy word lists confusing the selection and reducing the accuracy of the most accurate individual feature. Because of the reduction of accuracy this method of combining features was abandoned after the first limited set of tests.

8.3.2.3 Combining Distances

The third method of feature combination consists of combining the distances which have initially been calculated between an entire test word and each reference word for each separate feature. This method of combination was first discussed by Soong and Rosenberg(1988). For each test word, i, one distance, $d_{fi,j}$ is calculated with every reference word, j, for each feature, f, so that the total distance $Dist_{i,j}$ for the test word with a particular reference word, over all the features, N_f, is

$$Dist_{i,j} = \sum_{f=1}^{N_f} (\alpha_f \frac{d_{fi,j}}{\sum_{j=1}^{N_r} d_{fi,j}}), \tag{8.6}$$

where $\sum_{j=1}^{N_r} d_{fi,j}$ is the sum of distances calculated for that test word and each reference word j for all the reference words (N_r) for a given feature f_i. α_f is a weighting factor, such that each feature is weighted based on its recognition ability where

$$\alpha_f = \frac{\text{Feature Accuracy(\%)}}{100} \tag{8.7}$$

The distances for each feature are normalised by ($\sum_{j=1}^{N_r} d_{fi,j}$) and then added together. The normalisation of the calculated distances is necessary so that distances with different magnitudes can be meaningfully combined. The reference word, chosen as the correct word, is that which has the smallest distance $Dist_{i,j}$.

To ascertain this method's ability and sensitivity, a subset of features was chosen with a range of high and low individual accuracies. Individual and combined accuracies are plotted in Fig. 8.21. This method gives a significant decrease of error when features which have high individual accuracies were used, however when combining features with low and high recognition accuracies, a slight drop in accuracy from the higher individual accuracy was produced. This drop in accuracy shows this method to be sensitive to the combination of parameters with widely varying accuracies and this method should only be used with accurate features.

Examining the error decreases shows the largest decrease occurs with the speaker-dependent recognition results. A 68% error decrease (2.5% to 0.8%) occurs for the speaker-dependent female speaker result when combining cepstrals (with a projection distance) and RMS features. However, all other speaker errors decrease when combining results of CEP with Euclidean distance and the CEP with projection distance. The speaker-dependent male error dropped 21%, (15.9% to 12.5%), the speaker independent NZ male error dropped 1% (23.3% to 23.0%), the speaker-independent NZ female error dropped 6.4%, (12.5% to 11.7%), the speaker-independent USA male error dropped 14.1%, (23.3% to 20.0%) and the speaker-independent USA female error dropped 10%, (25.0% to 22.5%).

Other combinations also produce error decreases. The combination of CEP and ZCEP produced error reductions of up to 6%. The error reductions are greater for speaker-dependent recognition than speaker-independent. This could be because speaker-dependent accuracy results are higher for all features thus leading to less confusions when combining the features.

8.4 ACCENT EFFECTS

Because both American (US) and New Zealand (NZ) speakers are employed in this trial it was thought that it may be possible to examine differences due to accents for particular features. With the database recorded in separate environments, however,

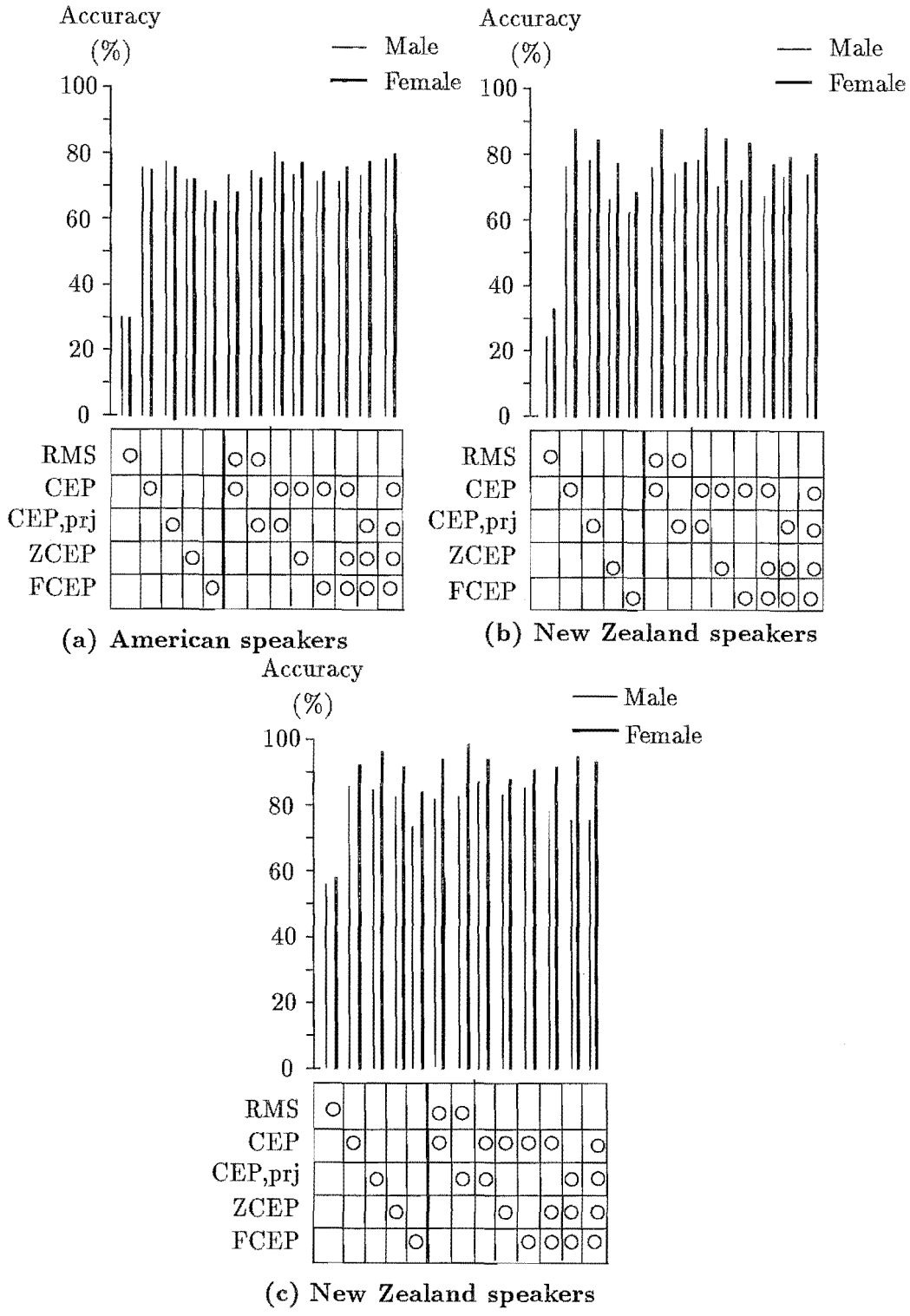


Figure 8.21. Recognition accuracies for a selection of features, and a combination of the features. (a) Recognition accuracies for American male and female and (b)New Zealand male and female speaker-independent tests. (c) Recognition accuracies for speaker-dependent tests with one New Zealand male and one New Zealand female. Recognition trials were carried out with pre-emphasis, Hamming windowing, and a UE21 DTW method with non-constant weightings. 10 reference templates per test were used, trained using the casual training method. Note CEP,prj = cepstral coefficient with projection distance.

it may be difficult to distinguish effects due to accent and effects due to recording artefacts. Many processing variables were standardised between the two databases including sampling rate and filter types. Other variables such as microphone type, background noise level, and background noise type could not be standardised. In the following discussion it is assumed that these variables are not so different between the databases as to cause excessively different recognition errors. This, however, may not be justified based on some recent findings by Kahn and Gnanadesikan, 1986 and also Baker and Pinto, 1986 which showed that up to 30% greater recognition errors were produced when variables such as microphone type and distance between microphone and speaker were changed. Despite these shortcomings an initial investigation of accent effects was carried out.

One particularly noticeable difference, shown in §8.3.1 (see Fig. 8.20), was that the American speakers have higher recognition accuracies with the ZCEP feature, while New Zealand speakers were better recognised with the CEP feature. It was also observed that PLPs gave up to 100% higher recognition accuracies for the American speakers over the New Zealand speakers. From these results there appears a definite relationship between feature type and accent.

Fig. 8.22 and Fig. 8.23 show the trends for each speaker-independent test with respect to the words tested. The figure shows the recognition accuracy for each individual word averaged for each speaker type; NZ male, NZ female, US male, and US female. The actual words recognised in these tests are given in the confusion tables, Tables B.1 through B.30 in appendix B The speaker-dependent confusion tables for one New Zealand male and one New Zealand female (given in Appendix B) are also given for comparison of word accuracies. Although these tables show the individual word errors for the test words it is difficult to ascertain from these tables any particular trend in word errors with respect to the speakers' accents or gender. Thus, although particular feature types do appear to depend on the accent of the user (PLPs were up to 100% better with American accented speakers than with New Zealand accented speakers), the individual word errors occur generally with the same words and therefore are not accent dependent. Fig. 8.22 and Fig. 8.23 also show that there is no trend in error for particular words between the US and NZ speakers. In fact there appears to be just as much variation between the male/female speakers of the same accent (NZ-NZ and US-US) as between male/male and female/female comparison of different accents (NZ-US). Thus it appears as difficult to recognise the speech of a speaker of the same accent but of different gender as it is to recognize the speech of a speaker of the same gender but different accent.

Examining Fig. 8.22 and Fig. 8.23 closely also shows the variation in relative accuracies for a particular feature or for a particular speaker. For example, for the word ZERO the NZ male has the highest accuracy when recognised with RMS. However, when recognizing the word ZERO using CEP the US male is recognised best (for a Euclidean distance) while the NZ female is recognised best with CEP (using a projection distance).

Another change in relative position is with the word that is recognised best between the features. For RMS the word SIX has the highest accuracy while for CEP (Euclid) the words ONE and NINE obtain the highest accuracies. For CEP (projection) the highest accuracy is with the words NINE and FOUR. As can be seen from Fig. 8.22 and Fig. 8.23 this change of relative position between words and speakers occurs for most of the words tested making it difficult to form trends or understand how particular database and accent parameters affect recognition accuracies. Also the change in relative accuracies for features dependent on speaker and vocabulary illustrates the difficulty of making general statements of recognition ability for features or systems

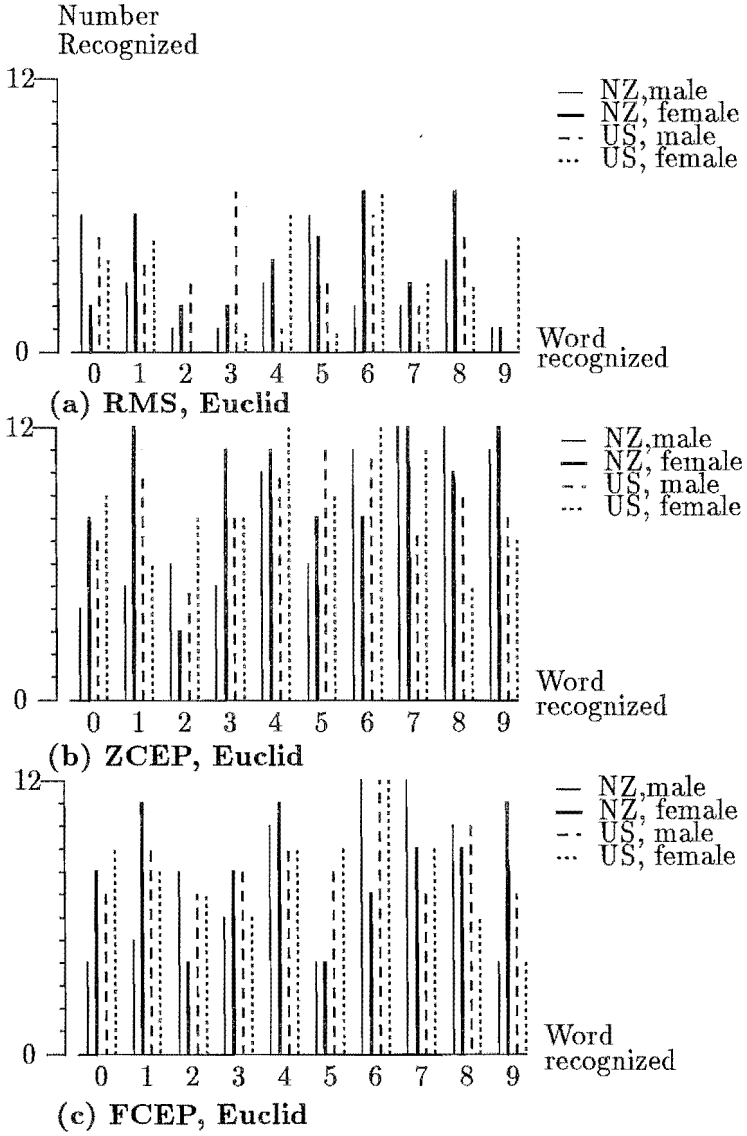


Figure 8.22. Recognition accuracies for a selection of features tested. A comparison of accuracies for 12 tests on each individual word for the features (a) RMS, (b) ZCEP, (c) FCEP. Results are for speaker-independent tests with American and New Zealand male and female speakers. Recognition trials were carried out with pre-emphasis, Hamming windowing, and a UE21 DTW method with non-constant weightings. 10 reference templates per test were used consisting of speech from the same accent and gender type and trained using the casual training method. The distance measure used in these trials was Euclidean.

from limited recognition tests.

With both accents giving similar word error rates it seemed interesting to test whether words of one accent could be used to train a system tested with the words of another accent. This test would ascertain how important it is to test and train a system with the same accent. A speaker-independent test was done in which the American and New Zealand data was both used. Either the American data was used as reference to test the New Zealand speakers or the New Zealand data was used as reference to test the American speakers. Ten reference and ten test templates were randomly chosen from the full American and New Zealand databases. The DTW UE21

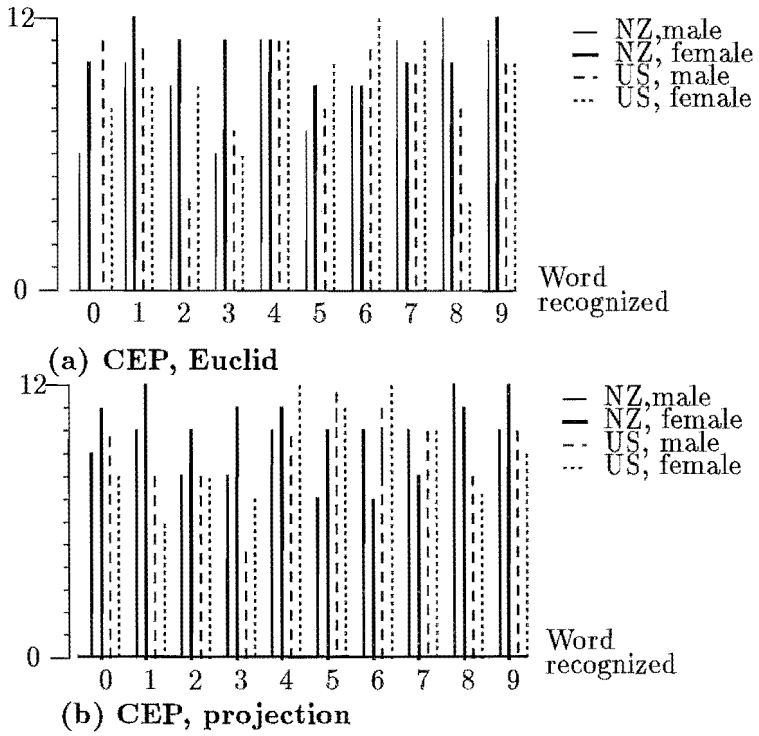


Figure 8.23. Recognition accuracies for a selection of features tested. A comparison of accuracies for 12 tests on each individual word for the features (a) CEP, (b) CEP with projection distance measure. Results are for speaker-independent tests with American and New Zealand male and female speakers. Recognition trials were carried out with pre-emphasis, Hamming windowing, and a UE21 DTW method with non-constant weightings. 10 reference templates per test were used consisting of speech from the same accent and gender type and trained using the casual training method. Note that the distance measure in these trials was either Euclidean or, for the graph labelled projection, cepstral projection distance measure.

method was used and a variety of features tested. The accuracies were extremely low and are given in Table 8.8. It is interesting to see from this table that the accuracies are substantially higher with the American speakers as reference (average female accuracies 30.0% (USA reference) and 14.9% (NZ reference), average male accuracies are 31.0% (USA reference) and 24.3% (NZ reference)).

When comparing American to American recognised speech with New Zealand to New Zealand recognised speech, as so far discussed in this chapter, there is no obvious trend or differences in the words recognised (refer Fig. 8.22 and Fig. 8.23) to point to any accent related differences. However, when examining the accuracies for American to New Zealand recognised speech and New Zealand to American recognised speech, as shown in Table 8.8, recognition accuracies are so low that it is obvious that accents must be affecting the accuracy. Although it is difficult to ascertain exactly what differences in accent affect the recognition accuracy, one difference which was particularly obvious between the two sets of words was the word length. The American speech database had on average 35% longer words than the New Zealand counterpart. Such differences could be accent related or database related hence showing the difficulties of using words recorded under different environments.

Sex	test	ref	RMS	ZX	LPC	CEP	CEP,proj	CEP,we	ZCEP	FCEP
Female	NZ	US	27	18	21	35	53	47	26	30
Male	NZ	US	16	15	25	45	55	50	35	26
Female	US	NZ	21	14	13	13	23	24	10	10
Male	US	NZ	13	20	24	26	35	18	26	26

Table 8.8. Percentage recognition accuracy when testing American speakers with New Zealand speakers as reference and testing New Zealand speakers with American speakers as reference. Results for both male and female speakers. Note that the distance measures used is Euclidean except for the columns labelled proj (projection distance) and we (quefrency weighted Euclid).

8.5 DISTANCE MEASURES AND ACCURACY

As discussed in Chapter 6 the distance measure can significantly affect the recognition accuracy. To test how great an affect the distance measure has on accuracy three types of distance measures were investigated based on the most accurate feature, cepstrals. The three measure were; unity weighted Euclidean, quefrency weighted Euclidean and cepstral projection. The mean accuracies and standard deviations from these tests are given in Table 8.9. Both speaker-independent and speaker-dependent recognition tests were performed on both male and female NZ and USA speakers. Jackknife tests were performed to obtain mean and standard deviation results from the tests. Results are tabulated for recognition accuracies using 2, 6, and 10 reference templates. All other recognition variables, such as windowing, pre-emphasis, DTW method etc were kept constant. Templates were selected by random selection (this method was used because, at the time of testing, the clustering method had not yet be implemented)

To test for significance in the results of Table 8.9 a paired-t test was used. From the t-test it was found that there was no significant difference (at the 95% level) between the weighted Euclidean and projection distance measures, however both the weighted Euclidean and the projection measure were significantly better than the Euclidean measure.

From the results of Table 8.9 it is impossible to say whether weighted Euclidean or projection is the better distance measure. It is, however, clear that both these distance measures perform well under the conditions in which they were used. Further testing is needed to ascertain their abilities in conditions with greater stress. Such conditions of greater stress would include noise, accent and vocabulary.

8.6 SUMMARY

For the purpose of building a real-time word recognition system a series of test have been undertaken and described. The processing techniques which give the most efficient recognition system, that is a system with the highest accuracy but also able to operate in real-time, were discussed. From the tests performed the optimum recognition system would employ a Hamming (or equivalent) window, a frame-size of around 200 samples, cepstral coefficients with either a projection or quefrency weighted Euclidean distance measure, and a dynamic time warping band or unconstrained endpoint method without feature combination. A clustering procedure should be used for training the recognition system to ensure the best reference template representation. The clustering procedure should produce 6 reference templates per word as this number of templates produced the best compromise between accuracy and storage requirements. Pre-emphasis and

Distance	2 templates/word	6 templates/word	10 templates/word.
Speaker-independent, male speakers, NZ			
Euclid	54.3 ± 4.3	68.2 ± 1.3	71.9 ± 2.3
WEuclid	57.7 ± 5.3	69.0 ± 1.2	70.0 ± 3.1
Projection	62.2 ± 2.5	69.2 ± 1.7	75.0 ± 2.2
Speaker-independent, female speakers, NZ			
Euclid	69.5 ±3.8	79.6 ± 3.4	86.3 ± 1.0
WEuclid	74.3 ± 4.3	85.4 ± 2.3	91.8 ± 1.4
Projection	67.9 ± 2.1	79.4 ± 1.4	81.5 ± 1.9
Speaker-independent, male speakers, US			
Euclid	60.8 ± 4.2	72.6 ± 1.3	74.3 ± 2.8
WEuclid	62.5 ± 4.5	78.5 ± 2.4	82.6 ± 2.3
Projection	61.4 ± 4.0	78.7 ± 2.3	82.6 ± 2.2
Speaker-independent, female speakers, US			
Euclid	62.3 ± 2.2	67.9 ± 1.6	72.4 ± 1.6
WEuclid	50.3 ± 5.5	67.5 ± 2.0	71.7 ± 2.3
Projection	66.3 ± 2.0	70.2 ± 1.3	75.8 ± 2.4
Speaker-dependent, male speakers, NZ			
Euclid	84.5 ± 0.7	90.7 ± 1.3	94.8 ± 0.5
WEuclid	83.0 ± 0.8	88.4 ± 1.3	91.4 ± 0.9
Projection	82.5 ± 0.9	89.0 ± 0.9	95.0 ± 0.8
Speaker-dependent, female speakers, NZ			
Euclid	88.8 ± 2.1	94.7 ± 0.8	97.5 ± 0.3
WEuclid	91.2 ± 0.7	97.1 ± 0.6	99.2 ± 1.2
Projection	88.6 ± 1.2	96.2 ± 0.6	97.7 ± 0.4

Table 8.9. Average speaker-dependent and speaker-independent recognition accuracies and standard deviation of the recognition accuracies for New Zealand and American accented speakers using cepstral coefficients with three different distance measures (Euclid, weighted Euclid and projection) and varying numbers of reference templates (2, 6, and 10). Tests were performed with UE2-1 DTW method with non-constant weightings which was trained using the casual training method. Averages were calculated by varying the reference and test templates using the jackknife method.

data frame overlap would not be used in such a system. For high operating accuracy the reference and test words must be taken from the same environment and from the same accent group.

From the tests in Chapter 8 speaker-independent accuracies as high as 92% were obtained and speaker-dependent accuracies as high as 100% were obtained. Generally, however, speaker-independent accuracies were around 79% and speaker-dependent accuracies were around 95%.

Chapter 9

AN EXAMINATION OF CONTINUOUS RECOGNITION WITH DTW

In this chapter real-time continuous speech recognition methods are examined which use the DTW methods previously discussed for isolated-word recognition.

It is the aim of computer recognition researchers to build computer recognition systems which are able to recognize unlimited continuous speech as a human listener does. To achieve this goal many continuous and connected speech recognition methods were researched in the early 1970s. Continuous recognition systems such as the DRAGON system (Baker, 1974), the SPEECHLIS system (Woods, 1975), and the HARPY system (Lowerre and Reddy, 1980) were initiated (refer Table A.3 in appendix A). Some of these systems are still being researched and developed today and now have very large vocabularies (between 1000 to 30000 words) and recognize complicated sentences with high accuracies. These systems recognize large vocabularies by relying on complicated methods of linguistic and semantic parsing to increase their word and sentence accuracies. Many of these systems also recognize smaller sub-units of the words such as phoneme, syllable, diphone or triphone segments (refer §2.1.3 and §2.2) requiring difficult and time-consuming segmentation procedures.

While the large vocabulary continuous word recognisers use complicated recognizing and parsing routines unsuitable for real-time operation, smaller vocabulary (10-500 words) recognisers often use the same routines as isolated word recognisers. Thus variations on DTW (refer §5.1) and HMM (refer §5.2) isolated word recognisers have been proposed to recognize sentences word by word (Rabiner and Levinson, 1985; Rabiner *et al.*, 1989; Godin and Lockwood, 1989; Myers *et al.*, 1981; Bridle *et al.*, 1982; Sakoe, 1979) and have been given the name of connected word recognition. Further, continuous speech DTW schemes which time align sentences and indefinitely long utterances have been documented (Chamberlain and Bridle, 1983; Bloom, 1984).

The advantage of connected word recognition systems is that they avoid the complicated problem of segmenting speech into smaller units thus by-passing inherent problems associated with segmentation. The fundamental problem with recognition involving segmentation is that any errors introduced by the segmentation cannot later be resolved by the recogniser and hence increase the overall recognition error of the system. It is generally considered that segmentation is difficult. Indeed some researchers believe it is impossible to say where one phoneme ends and the next one starts (Subrata, 1982) even though humans seem to be able to segment sounds very accurately by ear (Roach *et al.*, 1990).

This chapter examines two methods of continuous recognition for a small vocabulary and which can be implemented in real-time. These methods combine the ideas of continuous recognition, by using phonemes, and connected recognition by using DTW. The first part of this chapter examines a method which segments the utterances by reducing both reference and test words into phonemes, much like that proposed elsewhere for continuous methods (Woods, 1975; Lea *et al.*, 1975; Lowerre and Reddy, 1980). This

method will be known as the *phoneme segmentation* method. After the segmentation process this method uses a UE21 DTW procedure (refer §5.1), operating in real-time, to match phonemes for recognition. Overall recognition accuracy and the effects of segmentation errors on recognition accuracy are examined.

The second continuous recognition method uses a band DTW scheme to match reference phonemic representations to a whole test word (un-segmented). Because only the reference patterns are segmented, patterns can be processed prior to recognition and, to obtain high segmentation accuracy, may be processed by hand. Test words do not require segmentation and hence complicated automatic segmentation procedures are not required. To recognize the test word the one-pass DTW method is used, as described in §5.1.5.1. This method will be referred to as the *one-pass method.*

Although these methods are trialed on isolated words it is believed that both schemes can be extended to recognize sentences. It is considered that the extension to sentence recognition can be made because the system recognises individual phonemes and should be able to recognize sentence lengths of connected phonemes.

9.1 THE TESTS

Tests were carried out to assess the ability of a recognition system to recognize a number of vowel and consonant sounds. The recognition tests were performed with or without segmentation (for the phoneme and one-pass methods respectively). The aim of the tests was to establish the effect of segmentation accuracy on the recognition of the phonemes and to compare the abilities of the two types of recognition systems; the phoneme segmentation method and the one pass method. It is important to ascertain the significance of the segmentation error on recognition error because if the segmentation error produces insignificant recognition error then it would be easier to implement the phoneme segmentation method with coarse segmentation rather than the one-pass method. However, if segmentation errors cause large recognition errors then the one-pass method would be the better choice.

Both continuous word recognition methods were evaluated on a limited data set. A limited data set was employed because these tests were regarded as initial evaluations, devised to ascertain the relative capability of each method. A database consisting of only one male speaker with a vocabulary of the ten vowel sounds spoken in a /b/,/g/ context was used. Five repetitions of each word spoken by the male speaker were recorded. The vowels chosen are listed in Table 9.1.

A /b/,/g/ context was chosen to make segmentation relatively easy. The /b/ and /g/ sounds are produced with a closed mouth while the vowel sounds are produced with an open mouth. For this reason it was felt that there would be some acoustic cue which would distinguish the vowels from the consonants.

9.2 THE PHONEME SEGMENTATION METHOD

In this section the phoneme segmentation method is discussed. Recognition is preceded by both the reference and test words being automatically segmented into their individual phonemes. A DTW (refer §5.1) method is used to match the test phonemes to the reference phonemes for classification.

Rules for segmenting the phonemes were developed by examining a selection of features visually to isolate any that could be useful. The features examined were root-mean squared intensity (RMS), zero-crossing (ZX) rate, first predictor coefficient (LPC_1), second predictor coefficient (LPC_2), second cepstral coefficient (CEP_2), and the Euclidean distance between the first cepstral coefficient of adjacent frames, also known as the first

cepstral difference (ΔC_1). These features were extracted from frames of 100 samples which were Hamming windowed and pre-emphasised. No window overlap was used. A small sized frame was chosen because of the shortness of the consonant sounds. Many of the consonant sounds were only around 300 samples long (10kHz sampling rate) and it was important that the features extracted from the data represented the consonants accurately, that is not mixed in with data of the following or preceding vowel.

After visual examination of the six features it appeared that ZX was the most useful and this feature was used for segmentation. This feature might only be useful because of the limited test being undertaken, that is the small size vocabulary and the use of only one speaker. Because the aim of this investigation was not to design an accurate segmenter but rather to test the inter-dependence of recognition and segmentation accuracies the segmentation scheme was kept simple.

The ZX rate was used by setting an appropriate threshold, pictured in Fig. 9.1. The first rise of the ZX rate above the set threshold and then fall below the threshold indicated the beginning to ending of the first phoneme. The last rise above the threshold and a fall below the threshold indicated the beginning and ending of the third phoneme. The piece of the word lying in between these points was deemed to be vowel.

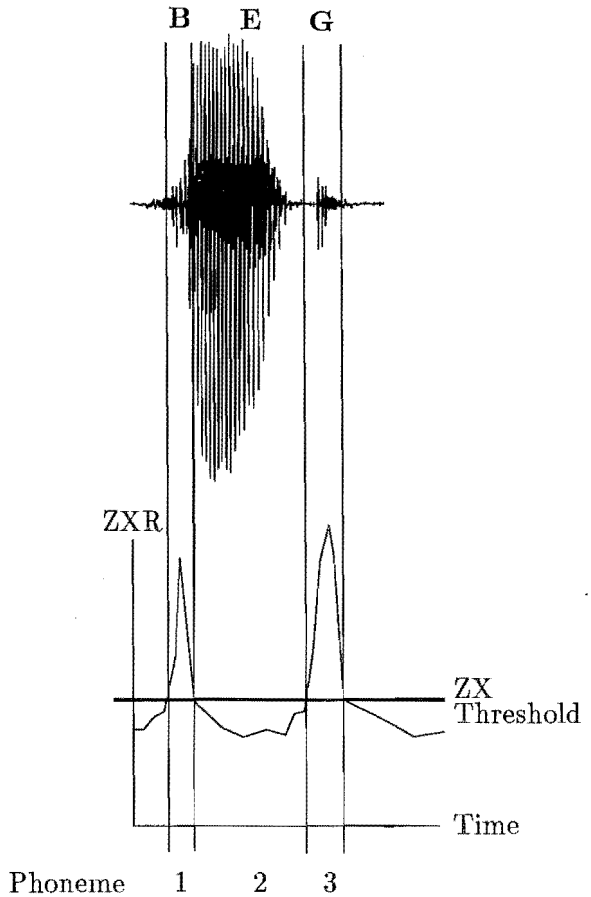


Figure 9.1. Endpoint determination by zero-crossings (ZX) calculations for the word BEG. Phoneme endpoints are found by setting a threshold on the ZX as shown.

Initially the threshold was set at an *optimum* value, that is a value which produced minimum segmentation error when tested with the trial word set. With this optimum value, and testing on the limited vocabulary, an average segmentation error of 22% was recorded. This estimated segmentation error rate was calculated by comparing

the automatic segmentation of the 50 words (5 tests for each of the ten words) against visual estimations of the endpoints from plots of the digitised speech waveform and the ZX waveform. A segmentation error was only noted when obvious errors of more than 2 frames were observed (this was recorded as one segmentation error). The segmentation error rate calculated for the 50 test words and for a particular level of ZX threshold was considered an estimate of the segmentation error rate for all words tested using that particular ZX threshold level. For the tests which investigated the effect of segmentation error the segmentation error was changed by changing the ZX threshold and examining segmentation placement. Any ZX threshold greater or less than the optimum threshold increased the estimated segmentation error.

WORD	Tested phoneme	Length (frames). Optimum threshold	Test1 Accuracy (%)	Length (frames). Non-optimum threshold	Test2 Accuracy (%)	Test3 Accuracy (%)
	/i/	27.5±1.5	100	29.2±0.86	100	44
bag	/æ/	30.2±2.1	100	28.8±3.8	80	100
barg	/ɑ/	32.2±4.5	80	35.6±1.0	100	100
beg	/e/	24.0±0.95	100	25.0±0.9	100	100
big	/I/	14.0±1.1	100	15.6±0.7	100	44
bug	/ʌ/	20.6±0.75	100	15.2±3.7	40	80
bog	/o/	26.4±3.3	100	22.6±4.6	80	100
boog	/U/	18.0±0.63	100	18.6±1.5	80	88
booog	/u/	37.2±5.6	100	22.4±5.8	80	100
boarg	/ɔ/	35.4±2.7	100	31.8±1.7	100	100
	/b/	3.8±0.2	96	3.8±0.2	88	93
	/g/	3.2±0.2	56	6.4±0.8	52	96
Vowel Accuracy (%)			98.0±2.0		86.0±6.0	85.6±7.3
Average Accuracy (%)			94.3±3.9		83.3±5.7	87.1±6.1

Table 9.1. Individual and average (with standard deviation) vowel and consonant recognition results from a segmentation based continuous word recogniser for one male speaker. Test 1 and Test 2 use 4 reference templates per word with reference and test templates jackknifed to produce 5 tests with unique data sets. Test 3 results are from a live recognition test via microphone into the real-time system. For Test 3, five reference templates are used per sound and are the same as those from Test 1 and Test 2. Test 1 and Test 3 use optimum ZX threshold giving estimated lowest segmentation errors (22%). Test 2 has the ZX threshold halved appropriate (optimum or non-optimum) ZX threshold set. A UE21 DTW method was used with producing an estimated 39% segmentation error. Mean and standard deviation of the length of the tested phoneme is also given, The mean and standard deviation of the length of each phoneme is calculated from a set of five test phonemes spoken by the single male speaker with appropriate (optimum or non-optimum) ZX threshold set.

Three sets of recognition tests were undertaken. The first and second tests jackknifed (refer §8.1.5) the five representations of each word between reference and test templates so that each of the five tests had a unique set of four reference templates. A cepstral projection distance measure (refer §6.3) was used in the UE21 DTW scheme. Test 1 differed from Test 2 by the threshold value used for segmentation. For Test 2

the segmentation threshold was half of that used for Test 1, increasing segmentation errors from 33 segmentation errors to 59 errors (39% segmentation error).

A third test was performed to ascertain the ability of the recognition method in a live situation. This test was undertaken to test if recognition performance would be severely degraded if there was a change in situation (such as background noise) while constants (such as segmentation thresholds) were kept the same. This test is presented as a preliminary trial only, obviously further testing is required. For the third test the speaker spoke directly into the recogniser via a Audio-Technica AT818II microphone (refer §8.1.3). Background noise at the time of the recordings varied but an average signal-to-noise (SNR) ratio was recorded at 22dB±10dB. Ten tests of each word were completed with each word having 5 reference templates, the same templates used in Tests 1 and Test 2. In the third test, as with the first test, the optimum ZX threshold was chosen giving estimated lowest segmentation errors. The recognition accuracy results of the three tests are given in Table 9.1.

Inspection of Table 9.1 shows, as expected, that recognition accuracy decreased when segmentation accuracy decreased with most recognition errors occurring when large segmentation errors occurred. Further, from an examination of the averages and standard deviations of the length of the phonemes and the recognition accuracies for each phoneme there appears to be a relationship between these variables. Recognition error increased when the segmentation error produced an increase in the standard deviation of the length or when the segmentation error produced shorter phonemes. Phoneme length decreases occurred for the vowels when the ZX threshold decreased because the vowels were deemed to lie between the consonants, which were found from the ZX threshold. When the ZX threshold decreased the consonants became longer hence decreasing the length of the vowels.

Increasing segmentation position error by 78.8% (22% segmentation error to 39% segmentation error) caused a 193% increase in overall recognition error (5.7% to 16.7%). The error increase was greater for vowels (which decreased in length), increasing the vowel error by 600% (2% to 14%) while only increasing the consonant error by 25% (24% to 30%). Further changes of ZX threshold produce changes in recognition error as shown in Fig. 9.2. The figure gives a plot of the calculated recognition error versus the estimated segmentation error, based on the 5 repetitions of the 10 words. This figure shows a monotonic relationship between increasing segmentation and recognition errors. To further quantify this relationship, however, further testing is needed.

9.3 THE ONE-PASS METHOD

For the one-pass method there is no segmentation of the incoming test word. Rather the test word is kept as a whole and compared with segmented reference templates. The reference templates are automatically segmented by the method discussed in §9.2, although they may be more accurately segmented by some other scheme prior to recognition if required. Reference templates are the same as those used in Test 1 and 2 in §9.2 and have the same length as that shown in Table 9.1 with optimum segmentation.

A one-pass (refer §5.1.5.1) band dynamic time warping (UEB) (refer §5.1.4) method is used for this continuous recognition system which used the weightings of §7.4.2. A band method is chosen because, as can be seen in Fig. 9.3, the endpoints for the comparison between reference and test are unknown prior to recognition. Not being able to place the endpoint, of the DTW warping path prior to recognition means that the boundary constraints required for a UE21 or CE21 DTW method cannot be set and hence a band method (UEB) must be used. With a band method the endpoints can be found during the recognition phase being dependent on the warping path.

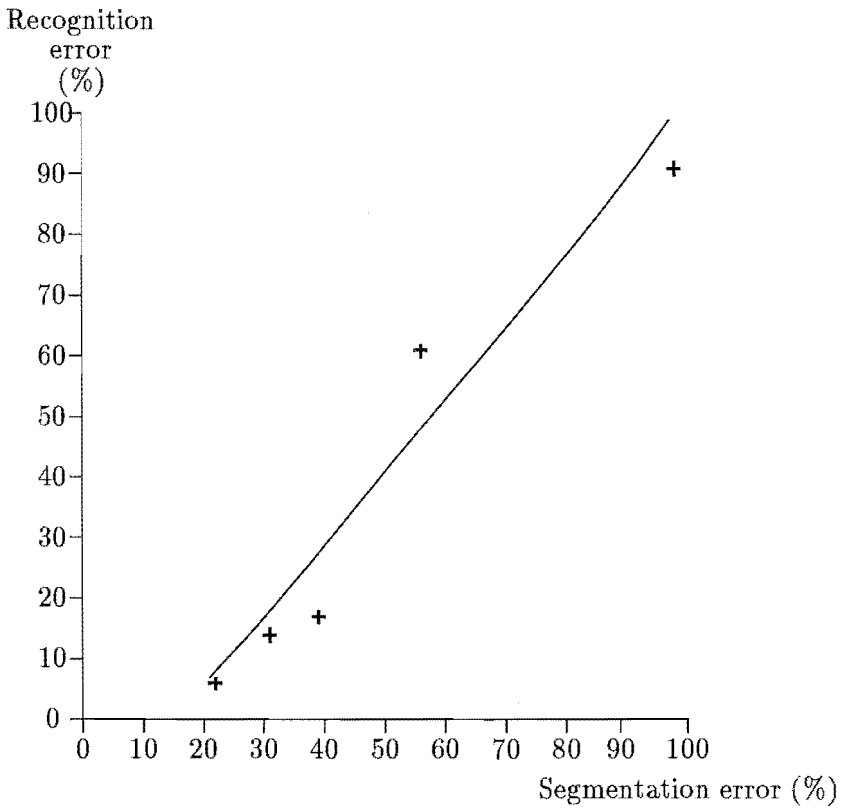


Figure 9.2. Plot of recognition error versus segmentation error for the phoneme-segmentation method. The plot shows a direct relationship between error in recognition and error in segmentation. Recognition tests were performed using a UE21 DTW method with cepstral projection distance, on one male speaker with 4 reference templates per word. Segmentation error was changed by changing ZX threshold value. Segmentation error was decided by visual examination of the data and segmentation position.

Using the UEB method the comparison of the test word with reference phonemes occurs as shown in Fig. 9.3. Firstly all reference phoneme templates are warped to the test word starting from the beginning of the test word. The reference phoneme with the smallest distance is deemed the best match and warping is continued at the end point of that warp, again using all reference templates. The endpoint for each warp is found by the warping procedure and is the optimum endpoint of that particular warping path segment. Warping continues until the end of the test word is reached. The selection of phonemes chosen to best match the word are those that produced the smallest distance at each warp.

Results for each phoneme are given in Table 9.2. From these results it is clear that the one-pass method gives lower accuracies than the phoneme recognition method with optimum segmentation. From Fig. 9.2 the one-pass method is as accurate as the segmentation methods with a segmentation error of approximately 40%.

The one-pass method is able to recognize vowel sounds with a much higher accuracy than the consonant sounds. One reason for this is that the consonant sounds are very much shorter than the vowel sounds and are swamped by the much longer vowels during recognition, because the distances calculated along the warping path are averaged together. Hence if large distances are calculated, say between vowel and consonant, for a very short time, the effect of these large distances is reduced because they are averaged with a large number of smaller distances calculated between vowel and vowel.

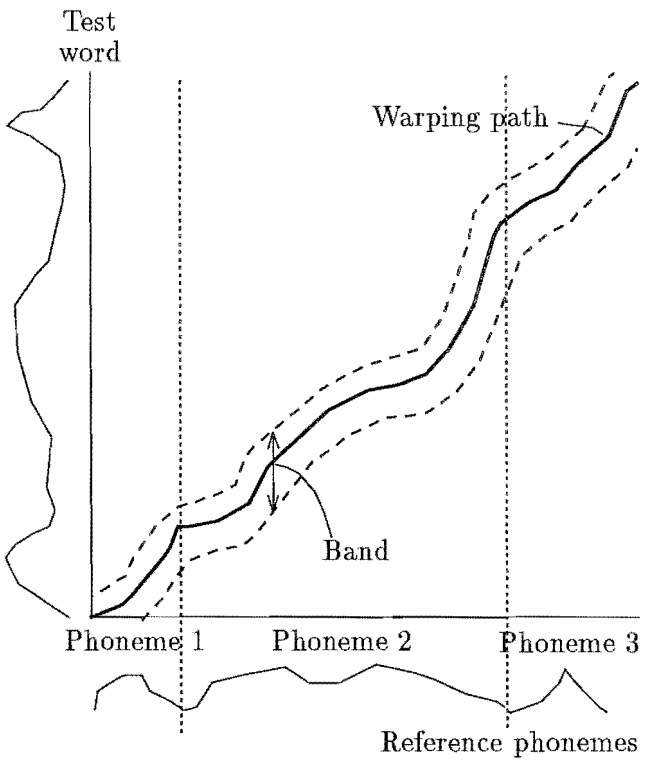


Figure 9.3. DTW unconstrained endpoint band method (UEB) used for continuous word recognition one-pass method. For the one-pass method the test word is not split into phonemes; rather recognition is achieved by matching all reference phonemes to the test word, starting at the beginning of the test word, and then finding the best reference template match as the reference phoneme with the smallest accumulated distance to its own endpoint. The next phoneme to be recognised is then found by continuing recognition at the end of the last recognised phoneme.

Thus the vowel to consonant distance has little effect on the overall distance calculated and the consonant is not recognised.

9.4 SUMMARY

Because the tests on continuous recognition are limited they are considered as initial comparisons only. It is clear, however, from this limited examination, that the segmentation accuracy is vitally important if a phoneme segmentation method of recognition is used. It is also shown that, at least for this vocabulary, if the segmentation is reasonably accurate (error <20%) then high accuracy can be expected from the DTW algorithm (>95%). Obviously, however, much further examination is required using many more speakers, vocabularies and conditions before an accurate error can be quoted.

It has also been shown that segmentation can introduce errors in the system which cannot be resolved by the recognition procedure thus increasing overall recognition error rate. Hence a method which does not require initial segmentation may be advantageous. The one-pass method was presented as such a recognition system. This method requires segmented reference phonemes while test words do not require segmentation prior to recognition. Rather, segmentation of the test word is performed during the recognition phase. Although the one-pass method produces lower accuracies than the segmentation

WORD	Test Phoneme	Accuracy (%)
beeg	/i/	40
bag	/æ/	100
barg	/ɑ/	80
beg	/e/	100
big	/I/	100
bug	/ʌ/	100
bog	/o/	100
boog	/U/	100
booog	/u/	80
boarg	/ɔ/	80
	/b/	10
	/g/	12
Average Vowel Accuracy (%)		88.0±6.1
Average Accuracy (%)		75.2±10.0

Table 9.2. Individual and average (with standard deviation) vowel and consonant recognition results from one-pass based continuous word recogniser. Segmentation of reference templates uses optimum ZX threshold. Tests use 4 reference templates per word and reference and test templates are jackknifed to produce 5 tests each with a unique data set. Testing is carried out on one male speaker, using a UEB method with cepstral projection measure.

method, this one-pass method does warrant further examination. One area of further examination for the one-pass method is the type of sub-word chosen. In the case reported phoneme sub-words were used, however difficulties with variation in the sizes of the sub-word units occurred. Accuracies may be improved if sub-word units such as demi-syllables were used which are of more constant length because each unit consists of both consonant and vowel.

Chapter 10

CONCLUSIONS AND SUGGESTIONS FOR FURTHER RESEARCH.

This thesis extensively reviews word recognition techniques, providing an exhaustive comparative study of many of the factors that affect recognition accuracy. From this comparative study information has been obtained to design and implement a real-time word recogniser of high accuracy. A real-time recognition system which operates at around 0.03 seconds per recognition has been built based on a TMS320C30. The real-time system achieves speaker-dependent accuracy greater than 95% and speaker-independent accuracy greater than 70%.

The experimental work, which has involved the testing of pre-processing variables, acoustic features, various DTW methods, various distance measures, various accented speech, and the testing of continuous recognition methods, is discussed in §10.1, and suggestions for future research are given in §10.2.

10.1 CONCLUSIONS

The aim of this research was to obtain the necessary knowledge and expertise to design and build a real-time word/speech recognition system. This has involved an exhaustive study of many factors which affect isolated-word recognition. To this end an extensive review of the speech and word recognition techniques discussed in the literature was undertaken. This review has prompted the many word recognition experiments presented in this thesis. The experiments have centred on four major areas of word recognition; pre-processing techniques, recognition features, recognition algorithms and distance measures. The trials of these techniques have been undertaken with New Zealand speakers in speaker-dependent mode and New Zealand and American speakers in speaker-independent mode, using the vocabulary of words ZERO through NINE.

10.1.1 Pre-processing Techniques

Results from pre-processing trials which included windowing, pre-emphasis, overlapping data frames, and varying data frame lengths were reported in §8.2.1. From these tests it is found that the most useful pre-processing technique is to window the data with a smooth function such as a Hamming, Hanning or Blackman window rather than a rectangular window, see §8.2.1.2. Tests with a Hamming window gave an average improvement, in recognition accuracy of approximately 20% over a rectangular window. Recognition accuracy increases are largest with linear prediction (LPC) and linear prediction derived cepstral coefficient (CEP) features, improving by 79.7% and 46.0% respectively. This is expected as LPCs model the speech spectral envelope more accurately when a smooth window is used (refer §8.2.1.2). Other features derived from LPCs, such as the transitional cepstral features, did not have a significant improvement. A lack of improvement for transitional features could be due to the averaging

over many frames of data which occurs with the calculation of this feature, removing the random noise introduced when using a rectangular window.

Pre-emphasis, discussed in §8.2.1.3, was the second most useful pre-processing technique. Pre-emphasis gave an average improvement of 6% across all speakers and features tested. Individual features, such as zero-crossing (ZX) and LPC, had significant performance improvements. While other features such as root-mean-squared energy (RMS) and perceptual linear predictor (PLP), had reductions in performance. Accuracy for the cepstrals (CEP) and transitional cepstrals (ZCEP,FCEP) representations was not significantly affected. The variations of accuracy for each feature and for each pre-processing techniques, such as pre-emphasis, shows how dependent a recognition system is on the feature chosen and pre-processing techniques used. Thus feature and pre-processing techniques must be carefully matched. Techniques such as overlapping data frames, §8.2.1.4, produced no significant improvement in accuracy for any features tested and would obviously not be incorporated into a recognition system.

Data processing parameters that significantly affected recognition accuracy were the size of the data frame (discussed in §8.2.1.1), and the number and choice of reference templates (discussed in §8.2.1.5). Recognition experiments revealed that the framesize from which features are extracted is at an optimum length at approximately 2 to 3 times a speaker's pitch period. Thus a frame length of 200 samples (at 10000 samples/s) appears an optimum choice for both male and female speakers. The choice of reference templates used for recognition can also severely affect the accuracy of the system. Recognition accuracies vary as much as 30% by changing the choice of reference templates of the same speaker. Such fluctuations in accuracy, due to simply changing the choice of reference templates, indicates the importance of correctly choosing a set of reference templates by some means, such as preliminary testing or clustering. Using a clustering technique can improve accuracy for a smaller set of reference templates (refer §8.2.1.5).

10.1.2 Feature Selection

Effects on recognition accuracies when using different features were examined in §8.3. The set of features considered covers a range of those often discussed in the literature. Both temporal representations such as zero-crossing rate and intensity (in the form of root-mean-squared, RMS), and frequency representation such as linear prediction coefficients (LPC), cepstral coefficients (CEP), zeroth order transitional cepstrals (ZCEP), first order transitional cepstrals (FCEP) and perceptual linear predictors (PLP) were examined. Speaker-dependent and speaker-independent trials were undertaken with both New Zealand and American speakers. For all speakers the highest recognition accuracies could be attributed to CEP and ZCEP features, whilst lowest accuracies were from RMS, ZX and PLPs. The relative usefulness of all these features were compared and it was found that relative accuracies were dependent on the speaker tested, the DTW constraints chosen, the number of reference templates and which reference templates were used. However CEP, ZCEP and FCEP features gave significantly better accuracies than all the other features over all speakers and methods.

10.1.3 Combining Features

An examination of three methods of combining features during and post recognition distance calculations to obtain higher recognition accuracy was reported in §8.3.2. The first method combined features during the DTW phase (§8.3.2.1), the second method combined the output word choices after DTW recognition (§8.3.2.2), and the third

method combined the calculated distances obtained for each feature from the DTW algorithm (§8.3.2.3). The first two methods gave no significant increase in accuracy or a slight accuracy reduction while the third method significantly improved accuracy under some limited conditions. The largest increase for the third method occurred with the speaker-dependent results giving up to 68% error reduction. Combination of features from speaker-dependent trials gave the greatest improvement because these results had the highest recognition accuracies for the individual features. With greatest individual feature accuracies combining results resulted in fewest confusions and hence greatest improvement. For speaker-independent tests, the third method had an average reduction of error of 10%. The error reduction was less for speaker-independent recognition than speaker-dependent recognition because the speaker-independent results were not as accurate. From these results it appears a combination of features is useful for highly accurate features and the method should be a combination of distances as of the third method discussed in §8.3.2.3.

10.1.4 DTW Methods

An examination of a selection of DTW methods was undertaken as discussed in Chapters 5 and 7. The methods tested were the constrained endpoint, 2-to-1 method (CE21), the unconstrained endpoint, 2-to-1 method (UE21) and the unconstrained endpoint band method (UEB). The UEB method was found to give the highest accuracy and to operate the fastest. The UE21 method, although not as accurate as the UEB method, was found to be significantly better than the CE21 method. Although the UE21 method's time per warp was longer than that of the UEB, the UE21 method's time can be reduced with the addition of a threshold as discussed in §5.1.6 and §7.7.

10.1.5 Accents Effects

New Zealand and American speakers were used to examine the effects of accent on recognition accuracy. It was found that when recognizing words from New Zealand speakers with a system trained with New Zealand speakers, or when recognizing words from American speakers with a system trained with American speakers, produced recognition errors which were not different between American and New Zealand speakers. Both New Zealand and American speakers had the same average accuracy across the words tested. No trends were found in actual word errors with both American and New Zealand speakers had similar word confusions (given in Appendix B). However it was found that some features did give differing results. The American speakers obtained a higher accuracy with the transitional representation (ZCEP) than the New Zealand speakers while the New Zealand speakers' accuracies were higher with cepstral (CEP) representation. It was also noted that the PLP accuracies were significantly higher for the American speakers than the New Zealand speakers. There was no testing, however, to establish whether these differences were dependent on accent or database (the NZ and US databases were recorded under different conditions) and thus no distinct accent effects could be attributed.

Further testing was undertaken with a mixed data set in which the American data was recognised from New Zealand reference data and New Zealand data was tested with American reference data. For the tests a series of features were examined, all giving low accuracies ($< 50\%$). Because accuracies for these tests were lower than that obtained from the speaker-independent tests with speakers of the same accent this suggests that accent does play an important role in recognizing one speaker's words from the words of another, different accented, speaker. Thus to obtain high accuracies it is important to train and test with speakers having the same accent.

10.1.6 Distance Measures

A selection of distance measures, which included Euclidean, weighted Euclidean and projection, were discussed in Chapter 6 and the results of tests were reported in §8.5. From the tests it is apparent that the choice of distance measure has a significant effect on the recognition accuracy. The best choice of distance measure was found to depend on the feature chosen to represent the speech and on the background noise of the speech. It should be noted that the background noise can have a large effect on the properties of the features used to distinguish the individual sounds and words (refer §6.2). From the results reported in §8.5 the cepstral projection and the quefrency weighted Euclidean distances were found to be significantly better than the Euclidean measure. Further testing with these two distance measures, including testing with noisy speech, would be required to more accurately quantify their abilities. At present, from the evaluation of this set of distance measures, the cepstral projection and the quefrency weighted distance measures perform equally.

10.1.7 Continuous Recognition

Two methods of continuous recognition using phonemes were examined in Chapter 9. The two methods, the phoneme segmentation method and the one-pass method, are based on the isolated word recognition methods discussed in §5.1.4 and tested in §7.6. Preliminary trials were run using these two methods to ascertain whether further research of these methods would be reasonable.

The phoneme-segmentation method segments the data into individual phonemes and recognises each phoneme separately with pre-segmented reference phonemes. For the 10 vowel and 2 consonant sounds tested an average accuracy of 90.7% was obtained (multiple test with a single male speaker). The accuracy of recognition for this method was found to be dependent on the accuracy of the segmentation. and this method therefore requires accurate segmentation (segmentation accuracy $> 80\%$). Accurate segmentation poses a large problem of its own with major controversy over the best approach to use, a method that does not use segmentation is appealing. The second method examined was a non-segmentation method using a one-pass DTW scheme. This method recognised whole words against previously segmented reference phonemes. The reference phonemes, which are saved before recognition, can be either hand-segmented or automatically segmented and checked, thus reducing segmentation error. Accuracy for the single male speaker, tested on 10 vowels and two consonants was 75% and hence lower than the phoneme method. Errors with the one-pass method were caused by the methods inability to cope with the variation in the lengths of the phonemes. Lengths of phonemes varied between consonant sounds, such as /b/ and /g/ which were as short as 2-4 frames long and vowel sounds as long as 20-30 frames long. Having such variation in the lengths of the phonemes means that distance calculated between short and long phonemes are averaged with distance calculated between long and long phonemes and hence error between short and long phonemes are averaged out. This averaging effect reduced the one-pass method consonant accuracy to as low as 11% while vowel recognition accuracy was 88%.

Further optimisation of the one-pass method is possible by implementing a multiple-pass scheme. This would incorporate properties of the two-level DTW scheme (refer §5.1.5.1) with the one-pass method. The idea would be to do multiple warps with the same phoneme, each time slightly changing the starting position. Multiple warping can occur over individual phonemes or over groups of phonemes. This would optimise the beginning and end points of each phoneme, taking into account coarticulation effects of complete word and sentence structures.

10.2 SUGGESTIONS FOR FURTHER WORK

Continuing research in this field should examine four areas. The first deals with distance measures for recognition since distance measures were found to significantly affect recognition accuracy. The second area for further examination is that of the effect of database parameters, such as noise and recording apparatus, on accuracy. The third topic to examine is the compensation recognition schemes require to cope with changes in speech accents and the fourth topic is to more thoroughly examine the uses of DTW for continuous recognition of small vocabularies. These four areas are discussed more fully in the following sections.

10.2.1 Distance Measures

The importance of the distance measure chosen for recognition, with respect to the accuracy of the recognition system, was discussed in Chapter 6. Because the choice of distance measure was found to significantly effect recognition accuracy it would be useful to test further a set of distance measures without being constrained to real-time considerations. In particular a wider selection of distance measures would be worthwhile examining whether distance measure weightings can be optimised for word recognition.

Also discussed in Chapter 6 was the effect of noise on the variation of acoustic features used to represent the signal and how this affects the choice of distance measure.

It would be worthwhile to examine the optimisation of a distance measure with respect to the noise level. This optimisation may result in a particular weighting, λ, added to the distance measure to equate noise levels of reference and test recordings such that

$$d_{opt} = \sum_{i=1}^{p} (c_{ref}(i) - \lambda c_{test}(i)), \tag{10.1}$$

as discussed by Mansour and Juang (1988) and in §6.2.3. For this distance measure the weighting function, λ, can be made dependent on background noise condition.

It may also be interesting to test whether optimising distance measures, by adding a weighting function, λ, can be used to reduce the differences between speakers of different accents. For such a case the weighting may be dependent on frequency, and be used to warp frequency components of the speech from speakers of different accents.

10.2.2 Database Parameters

Changing database parameters between training and testing can have a severe affect on the recognition ability of the system. One such effect of varying database and recording equipment is that of noise. Varying noise significantly degrades the accuracy of a recognition system and this area should be fully examined. The research in this thesis has avoided the topic of noise mainly because of the vast amount of research already undertaken in this field without any resolution of the noise problem. However it would be interesting to establish the affect of noise on the accuracy of this system, and to find which feature or features, and also which distance measure, may operate better under noisy conditions. Topics that could be examined could be the particular noise types (white, Gaussian, telephone etc) and noise levels. Also worth examining is the effect of varying the number of reference templates with different noise levels which may allow higher accuracies even under very low SNR levels. Other database parameters should also be examined, such as effects of varying frequency and phase response as would occur with varying recording devices, microphones etc.

10.2.3 Accents

Having access to both American and New Zealand speaker databases has allowed a tentative examination of how accent effects recognition accuracy. This examination has also shown how difficult it is to separate the accent affects from other database parameters which could affect accuracy. Although no conclusive results could be given it was noted that recognition accuracy was dependent on accent (particularly for PLP features). It would be interesting to examine more closely the effects of accent on accuracy, and in particular the effect on a particular features representing the speech. Tests undertaken to recognize one set of speech with a particular accent using reference speech having a different accent gave very low recognition accuracy. It would be useful for recognition systems if the problem of accent could be resolved so that high accuracies could be obtained when recognizing speech of speakers with different accents. Recognizing words of one accent using templates constructed from words of another accent would give an ultimate speaker independent recognition system. Further, reducing accent effects on recognition would allow reference templates to be saved only the once (usually during the production of the recognition system) and giving the user freedom from the laborious training task.

10.2.4 Continuous Recognition

A preliminary examination on the use of DTW as the word/phoneme comparator for a continuous recognition system was undertaken (Chapter 9). Two areas of further research could be carried out in this area examining both the phoneme method (§9.2) and the one-pass method (§9.3). The first is to look more fully into the area of segmentation and to examine accurate methods to divide words into the smaller units required for continuous recognition. The second is to examine fully the *one-pass* method. This method is appealing because it does not require automatic segmentation of the test word. Although lower accuracies were obtained for the one-pass method with the limited tests performed (§9.3) it is difficult to speculate whether the results obtained would hold for larger vocabularies and other speakers (both speaker-dependent and speaker-independent). The major fault of the one-pass method was the loss of accuracy when recognizing small sized phonemes occurring next to large sized phonemes. Possibly accuracies could be increased by using a better sub-word system, such as diphones or triphones. Such segments are of more equal lengths and may therefore give better accuracies.

Appendix A

RECOGNITION SYSTEMS THROUGH HISTORY

This appendix details, in tabular form, recognition attempts of this century. Discussed initially in Chapter 3, § 3.1, § 3.2 and § 3.3 the following tables give greater detail of the schemes, methods and implementations. Although I have tried to give a full overview of the systems that were produced, many recognition systems have been left out of the following tables leaving only those, I believe, have been most influential in the word and speech recognition arena.

The tables detail those points that I have considered most important for the comparison of different systems. These are; author, method, accuracy, vocabulary, speakers (tested and trained) and features. Where appropriate I have also included a short discussion of any other important points. These six major points do not, however, cover all aspects of these recognition systems, for further detail I direct the reader to the appropriate reference.

Table A.1. Early Recognition systems.

Author	Method	Accuracy	Vocab
KOPP,G.A., GREEN,H.C., 1946.	Spectrographic, distinguishing patterns.	-	All phonetic patterns.
POTTER,R.K., PETERSON,G.E., 1948.	Spectrographic, vowel representation. Time eliminated by tracing 'timeless' frequency path.	-	Vowels.
DREYFUS-GRAF,J., 1950.	Sonograph representation.	-	All speech.
POTTER,R.K., STEINBERG,J.C., 1950.	Specifying vowel sounds in terms of acoustical measurements.	Not sufficient information.	
SMITH,C.P.,1951.	Phoneme detection by comaparing frequency patterns.	-	Phonemes.
DAVIS,K.H., BIDDULPH,R., BALASHEK,S., 1952.	Digit recognition.	97-99%.	Digits "0" - "9".
OLSON,H.F., BELAR,H., 1956.	Phoneme recognising typewriter.	98%.	7 syllables
WIREN,J., STUBBS,H.L., 1956.	Phoneme recognition by successive-binary-selection	98-94%.	Vowels.
FRY,D.B., DENES,P.,1957.	Phoneme recognition: Sounds 72%. Words 44%.	12 English sounds, 140 English words.	
DUDLEY.H., BALASHEK,S., 1958.	Phoneme recognition.	>90%.	Digits "0"-"9".
DENES,P., 1959.	Phoneme recognition; Phonemes 60%. Words 24%. With linguistic information; Phonemes 72%. Words 44%. Non-trained speaker; Phoneme 45%.	13 phonemes.	
FORGIE,J.W., FORGIE,C.D., 1959.	Vowel recognition.	Without duration 88%. With duration 93%.	10 vowels, making 11 words.

Table A.1. Early Recognition systems.

Author	Speakers	Features	Discussion
KOPP,G.A., GREEN,H.C., 1946.		Spectrograms.	Focused on the legibility of visible spectrographic patterns.
POTTER,R.K., PETERSON,G.E., 1948.	Examined many speakers and variations of a single speaker.	Spectrograms, F1 versus F2 versus F3.	Designed to match the ear.
DREYFUS-GRAF,J., 1950.	All speakers.	Six frequency bands relating to the 6 principle frequencies of the mouth orchestra to decompose a word into its alphabetical elements.	Related the drawn patterns with the first known phoenician alphabet. The first attempt to write words from sounds.
POTTER,R.K., STEINBERG,J.C., 1950.	25 speakers. Male, female, and children.	Spectrographic.	Aural identification by 70 listeners. Two-dimensional representation by F1 versus F2 is not sufficient for all vowel discrimination.
SMITH,C.P.,1951	Single speaker.	Energy concentration in the frequency spectrum using filter bands 100-7000 Hz and the energy versus time.	
DAVIS,K.H., BIDDULPH,R., BALASHEK,S., 1952.	One male.	Zero-crossing rates from two channels (0-900Hz, 900-4000 Hz) giving rough F1 versus F2 plots. Matching by correlation.	-
OLSON,H.F., BELAR,H., 1956.	One male.	Quantizes into 8 frequency and 5 time intervals and one amplitude level.	
WIREN,J., STUBBS,H.L., 1956.	One male	Frequency bands.	
FRY,D.B., DENES,P.,1957.	One speaker.	Spectral content from 20 bandpass filters, 100 - 8000 Hz.	
DUDLEY.H., BALASHEK,S., 1958.	One male.	Frequency from ten bands, 0-3000Hz.	
DENES,P., 1959.	One speaker	Spectral content from 18 bandpass filters, 160-8000Hz.	Uses linguistic information about digram frequencies of phonemes.
FORGIE,J.W., FORGIE,C.D., 1959.	11 male, 10 female.	35 frequency channels (100-10000 Hz). Recognition based on F1 versus F2 locations plus durations.	

Table A.1. Early Recognition systems.

Author	Method	Accuracy	Vocab
DENES,P., MATHEWS,M.V., 1960.	Digit recognition.	Error: Speaker-dependent: no time normalization 13%, time normalization 6%. Speaker-independent: time normalization 30-40%.	Digits "0" - "9".
WELCH,P.D., WIMPRESS,R.S., 1961.	Vowel recognition by multivariate statistics.	Error: F1,F2 13%. F1,F2,F3 7-9%. F0,F1,F2,F3 6-7%. F0,F1,F2,F3,L1,L2 5-6%.	10 vowels.
OLSON,H.F., BELAR,H., 1961.	Phonetic reognition.		96 syllables.
SAKAI,T., DOSHITA,S., 1963.	Speech recognition by segmentation and classification.	Vowel 90%. Consonants 70 %.	Japanese monosyllables, consonant-vowel (CV).
REDDY,D.R., 1967.	Connected speech recognition by segmentation and classification.	Phoneme 81%.	30 seconds of continuous speech, 300 phonemes.
TEACHER,C.F., KELLETT,H.G., FOCHT,L.R., 1967.	Isolated digits.	90% with 1% mirecognition.	Digits "0" - "9".
GILLI,L., MEO,A.R., 1967/68.	Isolated digits.	Ten male speakers used also for training 100%. Other voices 70%.	Ten digits (Italian).
COMER,D.J., 1968.	Examined waveform asymmetry.	98%.	15 words.
PURTON,R.F., 1968	Isolated word, speaker dependent.	90%.	10 words.
SHEARME,J.N., LEACH,P.F., 1968.	Speaker independent word recognition. Multiple template representation, 1 per speaker to 10 per speaker.	1 template 60%. 10 templates 90%.	32 words, including digits.
BEZDEL,W., BRIDLE,J.S., 1969.	Word recognition by sound discrimination.	Digits 94.3%. Other words 98%. Average accuracy 96%.	14 words, including digits.
LAVINGTON,S.H., 1969.	Word recognition.	97%	Digits "0" - "9".
EWING,G.D., TAYLOR,J.F., 1969.	Digit recognition	'Excellent'	Digits "0" - "9".

Table A.1. Early Recognition systems.

Author	Speakers	Features	Discussion
DENES,P., MATHEWS,M.V., 1960.	Five speakers	Time versus frequency patterns. Reference patterns made from average many speakers' patterns. 17 channel filter bank, 200 - 4000 Hz. All words normalized to 60 time frames (linear normalization).	
WELCH,P.D., WIMPRESS,R.S., 1961.	10 male, 10 female.	Formant frequencies (F0 , F1, F2) and formant energy (L1, L2).	
OLSON,H.F., BELAR,H., 1961.		8 frequency bands (250-20000 Hz), quantized into 5 time steps.	
SAKAI,T., DOSHITA,S., 1963.	Male and female.	Formants and zero-crossings.	Speaker-independent recognition by adjusting for a particular voice.
REDDY,D.R., 1967.	One speaker.	0-10000 Hz, segmentation by intensity and zero-crossing, 21 parameters extracted for recognition, including formant values, energy, pitch and duration.	Operating speed at 40 times real-time.
TEACHER,C.F., KELLETT,H.G., FOCHT,L.R., 1967.	Ten male.	Single equivalent formant representation of the 3 formants.	
GILLI,L., MEO,A.R., 1967/68.	Ten male.	17 filter bands, 110 - 5600 Hz	
COMER,D.J., 1968.	Speaker trained	Waveform asymmetry.	
PURTON,R.F., 1968	One speaker.	Zero-crossings from two frequency bands, 200-1000Hz, 1000-4000 Hz.	
SHEARME,J.N., LEACH,P.F., 1968.	Ten male.	Normalized spectral envelope. Removal of timing information by setting constant word length.	
BEZDEL,W., BRIDLE,J.S., 1969.	30 speakers	Zero-crossing rates in full frequency band as well as high frequency and low frequency bands.	
LAVINGTON,S.H., 1969.	19 speakers	Number of zero-crossings and turn-arounds (zero gradient points) per 10ms.	
EWING,G.D., TAYLOR,J.F., 1969.	Five male.	Zero-crossing rate.	

Table A.2. Discrete Word Recognition systems of the 1970s

Author	Method	Accuracy	Vocab
RABINER,L.R., WILPON,J.G., 1979.	DTW, speaker independent, word recognition.	97% for 38 of the 40 speakers, 2 speakers gave 'poor' results.	54 computer word vocabulary.
TAPPART,C.C., DAS,S.K., 1978.	DTW, reduced search by reduction in samples saved.	90% for 40-50% memory reduction.	47 word alphabet + numbers "0" - "20".
RABINER,L.R., ROSENBERG,A.E., LEVINSON,S.E., 1978.	DTW, CE2-1 UE2-1 UELM	highest to lowest accuracy UE2-1 UELM CE2-1	10 words, consisting of digits, letters and words.
SAKOE,H., CHIBA,S., 1978	DTW, speaker dependent. Test carried out on : - symm/asymm - symm only - various systems taken from literature.	Errors; - asym 1.5 - 0.5, symm 0.3 - symm 1.9 - 0.9 Test from literature; Velichko and Zagoruko 2.7 - 2.0, White and Neely 1.3-0.33, Itakura 1.3-0.4, linear: 5.9-0.87	Japanese digits words and 50 Japanese geographical names.
LEVINSON,S.E., 1977.	Syntax used.	Errors reduced from 10% to 0.2%.	1000 randomly generated sentences, average length of 10.3 words from 127 word vocabulary.
LIN,W.G., CHAN,C.F., 1977.	Statistical pattern recognition. Semantics via energy and LPCs. Phoneme recognition using LPC, energy, zero crossings and first formant. Lexicon constraints.	Greater than 95% with training data	60 phonetically balanced words.

Table A.2. Discrete Word Recognition systems of the 1970s

Author	Speakers	Features	Discussion
RABINER,L.R., WILPON,J.G., 1979.	Sampling at 6.67kHz. Training USING 50 male and 50 female speakers. Testing using 40 representations, 10 from training set.	Training by clustering using autocorrelation. Testing with 8 pole LPC.	Online endpointing to remove spurious peaks. Fully automatic clustering techinque using K-nearest neighbour rule. Real-time operation on a CSP MAP-200 array processor.
TAPPART,C.C., DAS,S.K., 1978.	1 male, 4 repititions. 1 repitition for training, 3 repititons for testing.	12 bit PCM, 10kHz sampling, 240 point FFT, power spectrum every 12ms, 24ms Hamming window. Output in 6 bands.	Recognition accuracy drops as memory reduction increases. At 3% storage, accuracy is only 25%.
RABINER,L.R., ROSENBERG,A.E., LEVINSON,S.E., 1978	50 male, 50 female. 10 repititions per speaker.	Sampling at 6.67kHz. 8 pole LPC, Itakura distance.	Automatic endpointing, checked for errors.
SAKOE,H., CHIBA,S., 1978	Tests on two sets : - 10 males - 2 male, 2 females.	10 channel bandpass filters, Chebyshev norm distance.	Recognition requires 3 seconds per digit and 30 seconds per name.
LEVINSON,S.E., 1977.	All speaker	Finite state machine	Designed for airline flight and reservation information.
LIN,W.G., CHAN,C.F., 1977.		Sampling at 10kHz, 10 pole LPC, energy and zerocrossings. Pitch synchronous analysis for voiced speech.	Uses Baysian classifier for inital segmentation. Speech input in soundproof room.

Table A.2. Discrete Word Recognition systems of the 1970s

Author	Method	Accuracy	Vocab
GUPTA,V.N., GOWDY,J.N., BRYAN,J.K., 1977.	Comparison of distance measures, using dynamic and linear time warping. Speaker-independent.	- 6 autocorrelation coefficients, Euclidean 73.6%(DTW) 69.1% (linear). - 6 autocorrelation coefficients, log Euclidean squared 78.7%(DTW) 76%(linear). - 6 autocorrelation coefficients, Mahalanbois 80.4%(DTW) 75.7% (linear). - 6 autocorrrelation coefficients, cos(angle) 72.3%. - 14 LPCS, Itakura 71%(DTW), 82%(linear). - Autoreggressive distance measure 81.2%(DTW), 77.4%(linear).	40 words
LEVINSON,S.E., ROSENBERG,A.E., FLANAGAN,J.L., 1977 (Bell Labs).	Itakura DTW. Tested for speaker-dependent and speaker-independent. Input via telephone.	Error : 13 speakers 8.4% - Speaker-dependent 11.7%(word), 0.4%(word parsed), 3.9 %(sentence). - Speaker-independent (1 template/word) 45.5%(word), 5.6%(word parsed), 35.3%(sentence). - Speaker-independent (composite template/word) : 34.9% 6.5%(word parsed), 37.2%(sentence).	84 isolated words and 37 connectd word utterances.
WHITE,G.M., NEELY, R.B., 1976.	Comparison of parameters (LPC/BPF) and recognition methods (DTW/linear warping) based on Itakura(1975) log-likelihood distance.	BPF+DTW 98%(AN) 99.6%(words). LPC+linear 98%(AN) 90%(words). LPC+DTW 96%(AN)	North American state names (words) and alphanumeric (AN).

Table A.2. Discrete Word Recognition systems of the 1970s

Author	Speakers	Features	Discussion
GUPTA,V.N., GOWDY,J.N., BRYAN,J.K., 1977.	25 speakers.	Sampling at 10kHz, Hanning window. Autocorrelation and 14 pole LPC. Each utterance divided into 50 windows with 50% overlap.	Ten reference patterns per word. Jacknife procedure used during testing.
LEVINSON,S.E., ROSENBERG,A.E., FLANAGAN,J.L., 1977 (Bell Labs).	Speaker-dependent and speaker-independent.	LPC with Itakura distance.	Speaker able to repeat utterances before recognition to correct for speaking errors.
WHITE,G.M., NEELY, R.B., 1976.	Two males, five recordings each.	LPCs. Sampling at 10kHz, 25.6ms Hamming window, 50% overlap, 14 pole. BPF : filters span 10Hz to 10kHz, 20 1/3 octave bands, sampled 100Hz, smoothed and log scaled.	Claims to be the first published comparision of different methods. Utterances recorded in laboratory room, noise level at 65dB(A).

Table A.2. Discrete Word Recognition systems of the 1970s

Author	Method	Accuracy	Vocab
SAMBUR,M.R., RABINER,L.R., 1975.	Digit recognition, speaker independent. Tests with two noise conditions.	Error rate for the two environment; 2.7%. 5.6%.	Digits "0" - "9".
ITAKURA,F., 1975.	DTW using LPCs plus loglikelihood distance.	Tested with the two vocabularies; - 97.3%, 10 reps each word. - 88.6% 20 reps each word.	Tested with 200 Japanese geographical names, average number syllables = 3.5) and alphanumeric
DeMORI,R., 1973.	Time evolution of zero-crossing rate.	2% error.	Ten word digits.
ITAHASHI,S., MAKINO,S., KIDO,K., 1973.	Word dictionarys, (lexicon constraints), and phonological rules. Classification into phonemes.	- No dictionary 42-59% - Dictionary 92-93%	13 words.
ITCHIKAWA,A., NAKANO,Y., NAKATA,K., 1973.	DTW (fixed starting and free end points). Speaker-dependent. Testing sets of features.	- Spectrum 100%. - Cepstrum 100%. - Autocorrelation 92%. - LPC 77%. - Partial autocorrelation 98%.	Japanese digits.
POLS,L.C.W., 1971.	Recognition of vowels and isolated words.	- Vowels 98%. - Speaker-dependent words 95%. - Speaker-independent words 93%(male) 70%(female).	12 vowel sounds and 20 Dutch words including 10 digits.

Table A.2. Discrete Word Recognition systems of the 1970s

Author	Speakers	Features	Discussion
SAMBUR,M.R., RABINER,L.R., 1975.	For two environment tests: Five male and five female 30 female, 25 male.	Zerocrossings, energy of samples, normalized LPC energy, 2 pole LPC, sampling at 10kHz, frame size of 10ms.	Two tests consisting of - recorded in quiet room, high quality microphone, - noisy environment.
ITAKURA,F., 1975	One male.	Sampled (via telephone network) at 6.67kHz,200 sample Hamming window 50% overlap. 8 LPC coefficients. Noise level 68dB.	Introduced new distance measure for LPC all pole model representation. Run on a DDP-516 computer via telephone at 22 times real-time.
DeMORI,R., 1973.	Four male.	Two filter bands(LPF, HPF), giving zerocrossings. 7 output groups from low-pass filter, 4 output groups for high-pass filter.	
ITAHASHI,S., MAKINO,S., KIDO,K., 1973.	One speaker.	Sampled at 10kHz, Hamming window, frame size of 10ms, split into 4 frequency bands, linearly transformed.	Recorded in soundproof room.
ITCHIKAWA,A., NAKANO,Y., NAKATA,K., 1973.	One male, professional anouncer.	Features (dimensions); log spectrum (25), cepstrum (6), autocorrelation (4) LPC (7) PARCOR (5) Bandlimited to 4.2kHz. Sampled at 10kHz, 25.6ms frames every 10ms.	
POLS,L.C.W., 1971.	20 male.	17 and 18 1/3 octave filter bands, variance calclation to reduce dimensionality to 4.	1/3 octave filter bands used to simulate the ears hearing. Operates in real-time.

Table A.2. Discrete Word Recognition systems of the 1970s

Author	Method	Accuracy	Vocab
WARREN, 1971.	Fast pattern matching using tree-structure.	90-93%. Aprroximately 2 to 5% better than normal pattern matching.	Digits "0" - "9", 15 repititions for each speaker.
ITO,M.R., DONALDSON,R.W., 1971.	Zero crossings.	No definitive results.	Examined vowels, voiced and unvoiced fricatives and stops.
CLAPPER,G.L., (IBM), 1971.	Linear matching of frequency based word patterns. Speaker-dependent.	94.2%	Digits "0" - "9" plus 5 words (total of 16 words).
VonKELLER,T.G., 1970.	Digit recognition.	1-4% error and 1-4% rejection rate.	Training on 200 digits. Two tests: - 1000 digits recorded in soundproof room. - 500 digits in noisy room.
KLEIN, W., PLOM,R., POLS,L.C.W.,1970.	Speaker-independent, vowel recogniton.	4 dimensions 98%, 2 dimensions 88%.	600 vowels.
SCARR,W.A., 1970.	Speaker-independent digit recognition.	95% males, 89% females.	Digits "0" - "9".
VELICHKO,V.M., ZAGORUYKO,N.G., 1970.	Speaker-dependent word recognition by dynamic programming (DTW).	95%.	200 word.

Table A.2. Discrete Word Recognition systems of the 1970s

Author	Speakers	Features	Discussion
WARREN, 1971.	One male.	Infinitely clipped speech from 2 frequency bands, 0 to 1kHz and 1kHz to 4kHz, autocorrelated.	
ITO,M.R., DONALDSON,R.W., 1971.	One male.	Sampling at 16kHz. 2 frequency bands to 8kHz. Zero crossings.	Recorded in soundproof room.
CLAPPER,G.L., (IBM), 1971.	11 male and two female. 6 repititions giving 3360 utterances.	8 frequency bands up to 10kHz. Energy in each band. Word length cut into 6 time frames giving a 6 by 8 pattern for each word.	
VonKELLER,T.G., 1970.	Ten male for reference, four male for test.	Data segemented and each segment represented by F_1 and F_2 at beginning and end of frame pluse maximum F_1 during segment.	Recorded in noise free, sound proof room.
KLEIN, W., PLOM,R., POLS,L.C.W.,1970.	50 male.	18 frequency bands, reduced to 4 dimensions using principal component analysis. Recognition based on position of test word in the 4 dimensional space.	
SCARR,W.A., 1970.	Reference speech patterns from 30 male,30 female, tested by 12 male and 12 female.	Frequency range 100 to 7000Hz. Speech cut in to voice/unvoice, reduced to phonemes groups such as nasal, vowel, fricative, stop.	
VELICHKO,V.M., ZAGORUYKO,N.G., 1970.	Two speakers.	Energy output of five filters 0 to 8000Hz.	

Table A.2. Connected speech/word recognition systems of the 1970s

Author	Method	Accuracy	Vocab
SAKOE,H., 1979.	Two-level DP.	99.6%	Japanese digits.
SAKOE,H., CHIBA,S., 1979.	Two-level DP asymmetric.	Test on 4 vocabularies giving accuracies : 100% 99.6% 98% 100%	Four vocabularies tested : 200, 3 digit connected utterances. 50, 1-4 digit utterances. 1-4 digit utterances without word number specified. 50 geographical names combined to 3 word utterances.
MEDRESS,M.F., *et al*, 1977, Sperry Univac.	Sound class segmenter plus word sequencer to determine phrases.	Alphanumeric 95% (word), 88% (phrase). Commands 96% (word), 89% (phrase).	Alphanumeric and 64 commands.
WAKITA,H., KASUYA,H., 1977.	Vowel recognition in connected speech, speaker independent. Formant frequency scaling by normalization.	normalized formants 50.5%. unnormalized formants 42.2%.	
SAMBUR,M.R., RABINER,L.R., 1976.	Linear warping, statistical training.	Speaker-dependent 95-100%, average 97.6%. Speaker-independent 90-100%, average95.3%.	Connected digits, 3 digit strings 50, 3 digit strings. 20, 3 digit strings.
HATON,J., 1975.	DTW, plus syntactic (grammer) and semantic knowledge.	>99.9%.	36 french words.

Table A.2. Connected speech/word recognition systems of the 1970s

Author	Speakers	Features	Discussion
SAKOE,H., 1979.	Five male.	16 channel spectrum analyzer.	Recognition requires 10 seconds per digit. Dynamic programming used at word level. Bad segmentation removed by preprocessing. Speaker-independent recognition via speaker template adaptaion by using templates from each speaker for each word.
SAKOE,H., CHIBA,S., 1979.	Male speakers.	16 channel spectrum.	Recognition requires 10 seconds per digit. Important to remove 'intrinsically unreliable preliminary segmentation process'.
MEDRESS,M.F., *et al*, 1977, Sperry Univac.	Three male.	Sampling at 10kHz, 256pt FFT, off axis, 14 pole LPC to find F_1, F_2, F_3, F_0 (fundamental frequency) and energy from frequency bands.	8 year project.
WAKITA,H., KASUYA,H., 1977.	One male and one female.	Sampling at 10kHz, 14 pole LPC giving formants. 15ms frame size. Hamming window.	
SAMBUR,M.R., RABINER,L.R., 1976.	Speaker-dependent tested with three male and three female speakers. Speaker-independent tested with 10 male speakers.	Segmentation using voice/unvoice plus energy. Voiced regions analysed using 10 pole LPC, 10ms per frame. Training finds mean and variance of recogition parameters.	
HATON,J., 1975.	One male.	24 bandpass filters 110Hz to 7kHz, sampled every 10ms. Hamming metric distance.	Real-time

Table A.2. Connected speech/word recognition systems of the 1970s

Author	Method	Accuracy	Vocab
REDDY,D.Raj, ERMAN,L.D., NEELY,R.B., 1973, (HEARSAY).	Multiple knowledge sources; segmentation, aoustic recognition, syntactic knowledge, semantic knowledge,	Not given.	Chess moves.
NIEDERJOHN,R.J., THOMAS,I.B., 1973.	24 phonemes, connected English. Classification into voiced fricative, voiced non-fricative, unvoiced fricative, no speech. Word recognition from 24 phoneme classifications.	Classification 97%. Recognition 78%. Phoneme 87%.	Four 3-4 second sentences.
BARNETT,J., 1973, Vocal data management system (VDMS).	Multiple knowledge sources - syntax, semantics. Segmentation via acoustics. Non-linear parsing.	Not given.	1000 words.

Table A.2. Connected speech/word recognition systems of the 1970s

Author	Speakers	Features	Discussion
REDDY,D.Raj, ERMAN,L.D., NEELY,R.B., 1973, (HEARSAY).	Speaker-independent.	Five 1/3 octave bandpass filters, 200Hz to 6.4kHz. Output in intensity and zerocrossings in each band giving 12 parameters every 10ms.	
NIEDERJOHN,R.J., THOMAS,I.B., 1973.	One male.	Five filters up to 5kHz.	
BARNETT,J., 1973, Vocal data management system (VDMS).	-	Speech labelled	2-3 years still to run on a 5 year project.

Table A.2. Continuous speech recognition systems of the 1970s

Author	Method	Accuracy	Vocab
DE MORI,D., 1979.	Fuzzy sets and knowledge based.	94%	Vowel/consonant/vowel (VCV) words taken from continuous speech.
MEDRESS,M.F.,*et al*, 1978, (HARPY).	Knowledge based.		1011 words.
FUJISAKI,H., KUNISAKI,O., 1978.	/s/ and /sh/ discrimination using pole/zero model.	100% using poles and zeros, <100% with poles only.	60 Japanese words, VCV and CV.
JOHNSON,D.H., WEINSTEIN,C.J., 1978.	Speaker-independent phrase recognition. Phrases segmented into vowel/fricative/silence using probability density. Phrases recognized on syllable scoring.	Phrases 93-98%. Syllables 73-77%.	20 phrases.
COOK,C.C., SCHWARTZ,R.M., 1977. (Bolt, Bernak and Newmann) (HWIM)	Synthesis by rule to generate synthetic templates. DTW phoneme recognition.		1200 words.
ERMAN,L., 1977, (HEARSAYII).	Segmentation into classes, labelled into 98 different sounds using Itakura distance, parsing for sentence structure.	Word 80%. Sentence 90%.	1011 words.
ANDERSON,J.A., SILVERSTEIN,J.W., RITZ,S.A., 1977.	Vowel recognition by Neural network.	100% for all vowels after 10000 training sets.	Nine Dutch vowels.
SILVERMAN,H.F., DIXON,N.R., 1976, (IBM).	Recognition via spectra. Classification into broad classes and then phonemes.	Within Class 87.6%. Phoneme 64.6%.	33 phonemes.
ATAL,B.S., RABINER,L.R., 1976.	Classification via voice/unvoice/silence.	Errors; segmentation 5%.	Two test sentences and two training sentences.
HESS,W.J., 1976.	Pitch-synchronous phoneme classification on stationary parts only.	85.1%.	200 isolated words, 500 phonemes.

Table A.2. Conntinuous speech/word recognition systems of the 1970s

Author	Speakers	Features	Discussion
DE MORI,D., 1979.	Four male.	FFT and LPC formant tracking.	
MEDRESS,M.F.,*et al*, 1978, (HARPY).	Five Americans. Male and female.	LPCs	80 times real-time. Uses multiple sources of knowledge such as syntax, semantics. 5 year project funded at 3 million dollars per year.
FUJISAKI,H., KUNISAKI,O., 1978.	1 male.	Poles and zeros via mel scale FFT.	Zeros are found to be not important.
JOHNSON,D.H., WEIN-STEIN,C.J., 1978.	Six males for training. Ten males for testing.	Sampled at 6.67kHz. 10 pole LPC to calculate spectrum. 19.2ms Hanning window, shifted 9.6ms. 11 spectral values used.	2-3 seconds required per recognition.
COOK,C.C., SCHWARTZ,R.M., 1977. (Bolt, Bernak and Newmann) (HWIM)		Energy, timing, fundamental and formant frequencies. Frequency range 0 to 5kHz. Itakura distances.	
ERMAN,L., 1977 (HEARSAYII).		Sampled at 10kHz. Auto-correlation coeffiecients. Itakura metric.	
ANDERSON,J.A., SILVER-STEIN,J.W., RITZ,S.A., 1977.		Eight filter values.	
SILVERMAN,H.F., DIXON,N.R., 1976.	One speaker	Bandpassed to 8kHz. 20ms Hamming window, 80 point FFT.	5 times real-time. Speech recorded in soundproof room.
ATAL,B.S., RABINER,L.R., 1976.	One male for training. One male and one female for testing.	Sampled at 10kHz, frame size of 100 samples, no overlap. Zero-crossing, energy, adjacent speech sample correlation, first LPC from 12 pole LPC plus prediction error energy.	Recorded in anechoic chamber.
HESS,W.J., 1976.		Formants using peak picking, pitch synchronous segmentation via absolute signal amplitude.	

Table A.2. Continuous speech recognition systems of the 1970s

Author	Method	Accuracy	Vocab
WEINSTEIN,C.J., McCAND-LESS,S.S., MOND-SHEIN,L.F., ZUE,V.W., 1975, Lincoln Labs	Phoneme based. Prelim-inary segmentation into vowel-like, fricative-like and stop. Detailed clas-sification into dipthong, semivowel, nasal, vowel type, fricative type.	Error : Segmenation Vowel 1.4%. Fricative 1%. Stops 6%. Classification : Vowel 53%. Frictive 9%.	
LEA,W.A., MEDRESS,M.F., SKINNER,T.E.,1975.	Uses prosody to segment continuous speech and to locate stressed syllables for analysis.	Syntactic boundaries from fall-rise patterns of funda-mental : 90% accurate. Position of stressed sylla-bles from energy and fun-damental : 85%.	1000 seconds of continuous speech.
INGEMANN,F., MERMEL-STEIN,P., 1975.	Recognition by humans using spectrographic data. Human listener presented with spectrogram in three experiments.	Human recognition accu-racy - 75-83% limited. - 67% unlimited. - 48% monosyllabic.	
WALKER,D.E., 1975, Stan-ford research insti-tute (SRI system).	Uses multi-ple sources of knowledge, such as grammar and se-mantics, coordinated by a parser, to predict the se-ries of words.Classification into vowel, stop, voiced fricative, unvoiced frica-tive, silence and unknown.	44 out of the 77 utterances understood.	77 utterances.
DRUCKER,H., AND PREUSSE,J, 1975.	Recognition of ten vowel-like sounds.	Error <1%.	18 sentences, 13 vowel like phonemes.
JELINEK,F., BAHL,L.R., MER-CER,R.L., 1974, (IBM).	Linguistic statisti-cal decoder from phonetic symbols.	-	10000 words.
BAKER,J.K., 1974, (CMU).	Markov process us-ing probabalistic informa-tion and multiple sources of knowledge in an hierar-chical system.	100%.	Nine sentences.

Table A.2. Connected speech/word recognition systems of the 1970s

Author	Speakers	Features	Discussion
WEINSTEIN,C.J., McCANDLESS,S.S., MONDSHEIN,L.F., ZUE,V.W., 1975, Lincoln Labs		Spectrum analysis via LPC plus fundamental frequency.	
LEA,W.A., MEDRESS,M.F., SKINNER,T.E.,1975.	15 speakers.	Filter into 5 bands, up to 5kHz via hardware filters. Energy from bands.	
INGEMANN,F., MERMELSTEIN,P., 1975.	One female.	Spectrograms, 0 to 4.8kHz.	3 people used to recognise spectrograms. One experienced, one moderaely experienced, one with no experience.
WALKER,D.E., 1975, Stanford research institute (SRI system).		Sampled at 20kHz. Four digital filters up to 6.8kHz for voiced sounds, plus 3 filters up to 8kHz for unvoiced sounds. Sampled at 10ms. 28 pole LPC analysis of voiced intervals.	Recorded in soundproof booth.
DRUCKER,H., AND PREUSSE,J, 1975.	20 speakers.	Hardware circuit using analog threshold logic devices. Initially trained.	Real-time.
JELINEK,F., BAHL,L.R., MERCER,R.L., 1974, (IBM).	Speaker independent.	Input phonemes.	Language modelling and statistics.
BAKER,J.K., 1974, (CMU).	-	Knowledge sources such as - acoustic-phonetics, - lexical, - syntactic and semantic.	

Table A.3. Recognition systems of the 1980s

Author	Method	Accuracy	Vocab
DAVIS,S.B., MERMELSTEIN,P., 1980.	FFT and LPC based features examined; mel-frequency cepstrum coefficients(MFCC), linear frequency cepstrum coefficients(LFCC), linear prediction coefficients(LPC), reflection coefficients(RC), LPC derived cepstrum coefficients(CEP). Euclidean distance measure used for all representations except LPC. Log likelihood measure used with LPC. DTW recognition method.	Speaker 1 : MFCC 96.5%, LFCC 94.7%, LPCC 92.6%, LPC 85.2%, RC 83.1%. Speaker 2 : MFCC 95.0%, LFCC 87.6%, LPCC 87.3%, LPC 84.3%, RC 77.5%.	52 consonant-vowel-consonant (CVC) words.
RABINER,L.R., WILPON,J.G, 1981.	Distinguishing acoustically similar words by two-pass DTW approach. In the first pass the DTW recognizer provides a set of distance scores which are used to decide a set of possible classes in which the spoken word is estimated to belong. In the second pass a locally weighted distance is used to provide optimal separation within words from the chosen classes.	For the two test sets accuracy improvements of 6.6% and 3.5% were obtained.	Alphabet, digits and 3 command words.
MYERS,C.S., RABINER,L.R., ROSENBERG,A.E., 1981.	Examines two DTW methods, constrained endpoint band method and the non-constrained endpoint band method.	Non-constrained method consistantly better. String recognition rate of approximately 90%.	Tested on 54 computer words and also words embedded in continuous speech.
BROWN,M.K., RABINER,L.R., 1982	Uses both LPC and energy measurements in a DTW isolated word recognizer. LPC uses log likelihood ratio distance, energy using Euclidean distance.	LPC only; Male 90.7%, 87.6%. Female 77.5%,78.3%. Energy only; Male 22.5%, 27.9%. Female 24.8%,26.4%. LPC and energy; Male 96.9%, 89.9%. Female 83.0%, 83.0%.	Airline vocabulary, 129 words.

Table A.3. Recognition systems of the 1980s

Author	Speakers	Features	Discussion
DAVIS, S.B., MERMELSTEIN, P.,1980.	Two speakers.	Test on selection of features; mel-frequency cepstrum coefficients using 20 triangualr bandpass filters, linear frequency cepstrum coefficients from log magnitude discrete FFT, 10th order linear prediction coefficients from Hamming windowed data, relection coefficients derived from the LPCs and 10 LPC derived cepstrum coefficients. Signal filtered at 5kHz, sampled at 10kHz. Fourier spectrum calculated with 256 point Hamming windowed data frame.	Reference templates generated by averaged warped data.
RABINER,L.R., WILPON,J.G, 1981.	Two test sets each recorded over telephone line. Test set 1: 10 talkers not included in training set. Test set 2: 10 talkers included in training set.	-	The pair-wise word weighting curves were obtained by cross-comparing all word tokens within a word class, averaging the time aligned distance curves, and computing averages and standard deviations for each frame.
MYERS,C.S., RABINER,L.R., ROSENBERG,A.E., 1981.	Four speakers.	Recorded over telephone line, bandlimited to 3.2kHz, digitized at 6.67kHz, 45 ms of data represented by 8th order LPC. Distance scored calculated using Likelihood ratio.	
BROWN,M.K., RABINER,L.R., 1982	Two male and two female speakers.	LPC and energy.	Reference templates generated by clustering the speech of several male and female talkers to form six clustered templates of each of the 129 words.

Table A.3. Recognition systems of the 1980s

Author	Method	Accuracy	Vocab
TRIBOLET,J.M., RABINER,L.R., WILPON,J.G., 1982	Short-term and long-term features to represent speech in a weighted DTW based isolated word recognizer.	Speaker-dependent method improvement from 1 to 5.7%. Average accuracies; speaker1 95%, speaker2 95%, speaker3 92%, speaker4 87%.	39 word vocabulary of the digits, alphabet and command words.
RABIENR,L.R., ROSENBERG,A.E., WILPON,J.G., KEILIN,W.J., 1982.	Examines a 1109 word vocabulary LPC based DTW isolated word recognizer. Reduces complexity of recognition by constructeing smaller sized sub-sets of the vocabulary. Shows that by judicious choice of words for each sub-set can lead to significantly better recognition accuracies over random choice of word sets.	Approximate error rates for the various vocabulary subsets; talker1 2.5-20.8%, talker2 1.5-10.0%, talker3 9.0-29.5%, talker4 19.8-53.4%, talker5 8.0-30.9%, talker6 4.7-28.2%.	1109 words.
MERCIER,G., CALLEC,A., MONNE,J., QUERRE,M., 1982.	The KEAL recognition system, a continuous phoneme based recognizer.	Syllable segmentation accuracy of 95%, phonemic recognition rate of 61% giving word recognition accuracy of 93%.	Tested with 26 phonemic classes, isolated words (digits and commands) and continuous speech.
DAUTRICH,B.A., RABINER,L.R., MARTIN,T.B.,1983	Examines methods of filter-bank representation of speech for isolated word recognition.	Average error rates for male and female speakers; LPC 7.8%, best filter bank response (15 filters) 11.5%.	39 words consisting of alphabet, digits and words.
MOORE,R.K., RUSSELL,M.J.M TOMLINSON,M.J., 1983	Examines focusing the recognition decision during DTW at the point of the comparison that is able to discrimate similar sounding words. Uses the one-pass DTW method.	Average errors; Normal DTW 26.8%, discrimination method (small training set) 11.3%, discrimination method (large training set) 7.0%.	Six word pairs; stalagmite-stalactite, five-nine, rider-writer, bi-di, di-ti, and kei-dzei. Sixty representation of each word.
LEVINSON,S.E, RABINER,L.R., SONDHI,M.M., 1983.	Speaker independant isolated digit recognition based on LPC, HMM. A comparison is made with LPC, DTW isolated digit recognition.	Error rates; HMM 3.7%, DTW 3.5%.	Digits "0" - "9".

Table A.3. Recognition systems of the 1980s

Author	Speakers	Features	Discussion
TRIBOLET,J.M., RABINER,L.R., WILPON,J.G., 1982	Two male and two female.	Short-term analysis based on LPC calculation with short window size (15ms). Likelihood distance measure.	
RABINER,L.R., ROSENBERG,A.E., WILPON,J.G., KEILIN,W.J., 1982.	Speaker-trained with three male and three female.	LPC from speech recorded over telephone line. Filtered to 3.2kHz, digitized at 6.67kHz. Hamming windowed and preemphasised. 8-pole LPC .	
MERCIER,G., CALLEC,A., MONNE,J., QUERRE,M., 1982	Two male and two female.	14-channel vocoder giving a short-time spectrum every 13ms in the frequency range 250-4200 Hz. Speech is segmented into syllables and phonemes labelled. Consonant and vowels are recognized.	Method requires speaker adaption (speaker-dependent).
DAUTRICH,B.A., RABINER,L.R., MARTIN,T.B.,1983	Two male and two female.	13 different forms of filter banks were designed, filter banks used different number of filters, different filter shapes and different filter lengths. Results compared with LPC representation.	Speaker-independent by averaging multiple templates from various speakers.
MOORE,R.K., RUSSELL,M.J.M TOMLINSON,M.J., 1983	One speaker.	19 channel vocoder filter bank, output every 20ms. Euclidean distance.	
LEVINSON,S.E, RABINER,L.R., SONDHI,M.M., 1983.	50 male, 50 female. Same talkers for both training and testing but separate recordings.	LPC.	1000 utterance used to train HMM model.

Table A.3. Recognition systems of the 1980s

Author	Method	Accuracy	Vocab
BAHL,L.R., COLE,A.G., JELINEK,F., MERCER,R.L., NADAS,A., NAHAMOO,D., PICHENY,M.A., 1983.	A large vocabulary (5000 word) connected word recognizer.	94.5%.	5000 words.
ROSENBERG,A.E., RABINER,L.R., WILPON,J.G., KAHN,D., 1983.	Speaker-dependent isolated word recognizer based on demi-syllable (half-syllable) DTW recognition.	Demisyllable recognition error 18-33%, whole word recognition error 6-15%.	1109 basic English words.
HALTSONEN,S. 1984.	Examines constrained endpoint DTW and unconstrained endpoint DTW with the addition of extended test patterns and using reference patterns with silence frames added on the begining and ending of the word.	Error rates; Constrained endpoint 8.9%, Unconstrained endpoint 9.6%, Extended pattern (adding silence) 6.8%.	36 word alpha digit vocabulary.
RABINER,L.R., 1984	Addition of energy to the recognition of connected word sequences using a level-building DTW method and a level building HMM method. For one experiment syntax was also used. Both speaker-independent and speaker-dependent tests.	No significant improvement when adding energy to digit recognition. Improvements obtained when energy added for airline term vocabulary. String error rates; Speaker-dependent, airline vocbulary, DTW, no energy 20.6%, energy 13.2%, Speaker-independent, DTW, no energy 26.9%, energy 25.8%, HMM, energy 25.1%.	1520 connected digit strings plus 51 sentences taken from 129 word airline vocabulary
RABINER,L.R., WILPON,J.G., QUINN,A.M., TERRACE,S.G., 1984	Discusses a training procedure for a connected word recognition system which extracts training data from continuously spoken three word strings. Also used multiple reference patterns obtained by clustering.	String error rate 10.3% for carefully spoken sentences and 7.4% for normally spoken digit strings. For carefully spoken sentences embedded training decreased performance. For normally spoken strings a improvement of performance for embedded string training from 25.3% error to 15.3% error.	40 digit strings.

Table A.3. Recognition systems of the 1980s

Author	Speakers	Features	Discussion
BAHL,L.R., COLE,A.G., JELINEK,F., MERCER,R.L., NADAS,A., NAHAMOO,D., PICHENY,M.A., 1983.	Four male and two female.	Sampled 20kHz, energy from 80 frequency bands plus total energy.	Tested on 20 sentences covering 299 words.
ROSENBERG,A.E., RABINER,L.R., WILPON,J.G., KAHN,D., 1983.	Two male.	8-order LPC analysis from 6.67kHz bandlimited speech for every 45ms, shifted 15ms.	
HALTSONEN.S., 1984.	One male.	Speech digitized at 20kHz, 14 pole LPC extracted every 10ms from 5kHz bandpassed, preemphasised speech which is Hamming windowed.	One representation as reference, nine representations for testing.
RABINER,L.R., 1984	19 talkers for test set 1, 6 talkers for test set 2.	Telephone line speech recording.	For speaker independent tests tokens of 100 different talkers (50 male and 50 female) were clustered.
RABINER,L.R., WILPON,J.G., QUINN,A.M., TERRACE,S.G., 1984	19 speakers.	Recorded over telephone line.	Level build DTW method used.

Table A.3. Recognition systems of the 1980s

Author	Method	Accuracy	Vocab
LAU,Y., CHAN,C., 1985	Isolated digit recognition system based on zero crossings and energy.	97.2%.	50 utterances of each digit.
HALTSONEN, S., 1985	Level based dynamic time warping (DTW) for recognizing connected word sentences.	Error rate : clustering reduced error from 7.7% to 4.9%.	2000 word office correspondence.
FURUI,S., 1986.	Isolated word recognition system based on the combination of instantaneous and dynamic features of the speech spectrum.	Combining features gives error rate of 2.4%, instantaneous features only gives error rate of 6.2%.	100 Japanese city names.
HERMANSKY, H., 1987.	An auditory model of speech perception (PLP) is used as a front-end to a recognition system for speech and word recognition.	Speech recognition, PLP outperformed LPC measures. Isolated word task, speaker-dependent PLP outperformed LPC representation, speaker-independent PLP outperfomed LPC representation up to model order of 5.	Speech, 104-word typewriter key-board vocabulary, hand-labelled for phonemic events. Isolated word task involved 36 word alpha-digit.
BUSH,M.A., KOPEC,G.E., 1987.	Speaker independent connected digit recognition using acoustic-phonetic features. digit vocabulary is modeled using a finite state pronunciation netork.	String recognition accuracies of 96-97%.	Texas-instruments (TI) multidialect, connected digit database.
JUANG,B., RABINER,L.R., WILPON,J.G., 1987.	Bandpass liftering of cepstral coefficient representation for isolated vowel and digit recognition.	Digit error rate of 1% for speaker-independent recogniton(50% error reduction on LPC/likelihood distance).	4 sets of digits and vowels.

Table A.3. Recognition systems of the 1980s

Author	Speakers	Features	Discussion
LAU,Y., CHAN,C., 1985	Two male and three female.	Sampling at 6.67kHz, time frame of 7.5ms energy and zero crossings extracted. Distributions of the two measures were formed to distinguish the digit words.	
HALTSONEN, S., 1985	One male.	14 pole LPC on 0-5kHz speech sampled at 20kHz. Hamming window used. Likelihood distance measure used.	Clustering of training material by averaging. Patterns are extended by silence at beginning and ending. Four representations used for training.
FURUI,S., 1986.	Test set 1: 100 words uttered twice by four male speakers. The first utterance was used to train, the second utterance was used to test. Test set 2: 20 male speakers were tested with a trained system using first utterances of test set 1.	Instantaneous and dynamic cepstral coefficients. Features calculated from speech filtered to 4kHz and sampled at 8kHz. Hamming window of 32ms width used every 8ms.	Staggared array DP matching algorithm.
HERMANSKY, H., 1987.	Speech recogniton two male and two female speakers. Isolated word task, speaker-dependent and speaker-independent tests.	PLP and LPC calculated with various pole order.	Each vocabulary word represented by single template.
BUSH,M.A., KOPEC,G.E., 1987.	Speaker-independent.	11 acoustic attributes: short term LPC spectra; total energy and formant frequency contours; peak low-frequency energy; segment duration.	
JUANG,B., RABINER,L.R., WILPON,J.G., 1987.	50male and 50 female.	Raised sine liftered cepstral coefficients.	

Table A.3. Recognition systems of the 1980s

Author	Method	Accuracy	Vocab
MANSOUR,D., JUANG,H., 1988.	Discusses a set of similarity measures, called projection distance measure, for robust speech recognitition.	Recognition improvement under extreme noise conditions (SNR=5dB) accuracy improvement from 59% to 77.8%.	39 word alpha digit.
LEE,C., JUANG,B., SOONG,F., RABINER,L., 1989.	Comparison of three types of fundamental sub-word units: whole word, phoneme-like and acoustic segment.	Highest accuracy(95.1%) occurred with whole word units, then acoustic segment units(89.3%) and then phoneme units (81.6%).	1109 word.
BAHL,L.R., *et al.*1989.	The IBM continuous, large vocabulary recognizer. The recognizer consists of an acoustic processor, an acoustic channel model, a language model and a linguistic decoder.	Speaker-dependent recognition accuracy of 89% for continuous speech.	50 sentences drawn from spontaneously generated memos covered by a 5000 word vocabulary.
ZUE,V., GLASS,J., PHILLIPS,M., SENEFF,S., 1989	A phonetically based spoken language understanding system called SUMMIT. From the input speech signal a network of phonetic labels are found with scores indicating the system's confidence in the segmentation and labelling accuracy.	Speaker-dependent 77% accuracy. Speaker independent 70% accuracy.	Speaker-dependent, trained on 500 sentences, tested on 210 sentences. Speaker-independent, trained on 1500 sentences, 300 speakers, tested on 225 sentences from 45 speakers.

Table A.3. Recognition systems of the 1980s

Author	Speakers	Features	Discussion
MANSOUR,D., JUANG,H., 1988.	Two male and two female speakers.	Recorded over telephone line. Filtered at 3.2kHz, sampled at 6.67kHz. Manually endpointed, and analyzed to extracte 8 LPCs from 130 sampled frames overlapped by 80 samples. Rectangular window without preemphasis. 12 cepstral coefficients were extracted.	
LEE,C., JUANG,B., SOONG,F., RABINER,L., 1989.	Three male speakers.	Telephone line recorded speech, filtered to 3.2kHz, sampled at 6.67kHz, preemphasised and Hamming windowed. Nine autocorrelation values extracted.	Hidden Markov modelling for recognition.
BAHL,L.R., *et al.*1989.	Ten male talkers.	20 spectral features based on an auditory model, Euclidean distance.	Used hidden Markov modelling for word recognition.
ZUE,V., GLASS,J., PHILLIPS,M., SENEFF,S., 1989	100 speakers.	Auditory modelling used to perform acoustic segmentation.	

Table A.3. Recognition systems of the 1980s

Author	Method	Accuracy	Vocab
JUNQUA,J., WAKITA,H., 1989	Examines PLP and cepstral features. Cepstrals tested with liftering. Noisy conditions analyzed.	Database1, speaker-dependent, tested with Euclidean and projection distance measures. Cepstral features. Highest accuracy (93.33%) for noiseless case. Cepstral projection distance gave highest accuracy (66.67%) between non-noisy reference and noisy test. Also tested speaker-dependent with LPC, PLP and SLP (where SLP is the authors' method of auditory front end analysis). For clean/clean (reference/test) data LPC gave highest accuracy (Euclidean 94.4%, projection 94.1%), clean/noise SLP highest accuracy (Euclidean 54.73%, projection 60.28%),	Two databases tested. Database1 : alpha-digit, recorded in quiet environment. Two repititions. Database2 : 49 word consisting of alpha-digit and words. Recorded in quiet environment but speakers were listening to noise through headphones. Two repititions. White Gaussian noise added to both databases for testing.
HUNT,M.J., LEFEBVRE,C., 1989.	Comparison of several acoustic representations for speaker-dependent and independent connected and isolated-word recognition tests with undegraded and degraded speech (addition of white Guassian noise and spectral tilt). Acoustic representations include auditory model, cepsturm coefficients (derived from mel-scaled FFT), dynamic and instantaneous cepstrum coefficients, and sets of linear discriminant functions derived from filter-bank outputs.	Highest accuracies were from the linear discriminant functions derived from filter-bank outputs.	Isolated digits and connected digits.
LEE,K., HON,H., HWANG,M., MAHAJAN,S., REDDY,R., 1989.	A large-vocabulary speaker-independent continuous speech recognition system (SPHINX).	With grammer 96%. Without grammer 82%.	997 word DARPA resource management database.

Table A.3. Recognition systems of the 1980s

Author	Speakers	Features	Discussion
JUNQUA,J., WAKITA,H., 1989	Two sets of speakers; Database1: six male and four female. Database2: five male and five female. One template/word.	Samples at 10kHz. all-pole representations; LPC and PLP.	Generally found PLP with RPS front end generally better under noise conditions.
HUNT,M.J., LEFEBVRE,C., 1989.	Two database tested; Database1: Isolated words, 1346 digits from 9 male speakers and 900 digits from 5 female speakers. Database2: Connected-digits 1352 digits spoken by male speakers and 900 digits spoken by female speakers.	Mel-cepstrum, auditory modelling, filter-bank output, transitional cepstrum. Frame width of 6.4ms, three frame averaged to give output every 19.2ms.	DP matching.
LEE,K., HON,H., HWANG,M., MAHAJAN,S., REDDY,R., 1989.	Training on 4200 sentences spoken by 105 speakers. Evaluated on 12 speakers with 25 sentences per speaker.	Sampled at 16kHz, pre-emphasised, Hamming windowed (20ms) every 10ms. 14 LPCs are derived from which 12 cepstral coefficients are calculated, and transformed to	HMM modelling to identify 48 phonemes. Also tested on triphones.

Table A.3. Recognition systems of the 1980s

Author	Method	Accuracy	Vocab
WEINTRAUB,M., MURVEIT,H., COHEN,M., PRICE,P., BERNSTEIN,J., BALDWIN,G., BELL,D., 1989.	Speaker-independent speech recognizer developed at SRI, (DECIPHER). Recogntion by speech knowledge and linguistic concepts. Word models are formed from network representations of word pronunciations and from a set of phonetic models.		Sentence read in sound-isolated room, training on 3950 sentences from 105 talkers which are not part of the test set.
PAUL, D.B., 1989.	Speaker-independent continuous speech, large vocabulary recognizer developed at Linclon laboratories(MIT).	Speaker-dependent 96.5%, Speaker-independent 87.4%.	DARPA resource management task vocabulary, consisting of 991 words.
RABINER,I.R., WILPON,J.G., SOONG,F.K., 1989.	Evaluation of a connected digit recognition system using both instantaneous and dynamic cepstral coefficeints	Tests on speaker-dependent(SD), multi-speaker(MS) trained and speaker-independent(SI) recognition. For unknown string length error rates SD 0.78%, MS 2.85%, SI 2.94%.	Connect digit strings, up to 7 digits.
SVENDSEN,T., PALIWAL,K.K., HARBORG,E., HUSÒY,P.O., 1989.	Speaker-dependent speech recognition based on acoustic sub-word units.	Three distortion measures tested; likelihood ratio (10 LP coefficients) 84.7%, cespstrum Euclidean (14 coefficients) 77.3%, liftered cepstrum (14 coefficients) 90.0%. Compariosn between methods; whole word HMM:91%, whole word DTW:97%, Acoustic subword units : 90.8%.'	100 repititions of 42 Norwegian words. 50 utterance for training and 50 utterances for testing.

Table A.3. Recognition systems of the 1980s

Author	Speakers	Features	Discussion
WEINTRAUB,M., MURVEIT,H., COHEN,M., PRICE,P., BERNSTEIN,J., BALDWIN,G., BELL,D., 1989.	112 male and 48 female talkers.	Accuracies were from 74.1% to 93.7% with various lexicons.	Recognition testing on standard DARPA database. Based on HMM modelling.
PAUL,D.B., 1989.	Two modes; Speaker-dependent using 12 speakers. 600 training sentences, and 100 test sentences per speaker. Speaker-independent using 12 speakers. 2,880 training sentences from 72 speakers, 100 test sentences per speaker.	Mel-cepstrum.	Uses HMM modelling triphones.
RABINER, L.R., WILPON,J.G., SOONG,F.K., 1989.	Texas instuments, 224 talker standard database.	Instantaneous and dynamic cepstral coefficients. Data sampled at 6.67kHz, pre-emphasised and Hamming windowed in 45ms frames with 15ms spacing. 8th order LPC analysis is performed, and 12 cepstral coefficients derived. A sine based lifter is used. Dynamic cepstrals are calculated from the weighted cepstral coefficients over a 5 frame window.	Level building HMM modelling of connected digit words.
SVENDSEN,T., PALIWAL,K.K., HARBORG,E., HUSÒY,P.O., 1989.	One male.	Sampled at 8kHz, frequency cut-off at 3.5kHz. Sampled speech is pre-emphasized and 11 autocorrelation coefficients extracted every 15ms using a Hamming window of size 45ms. 10 LPC coefficients calculated from which 14 cepstral coefficients are found.	HMM modelling.

Table A.3. Recognition systems of the 1980s

Author	Method	Accuracy	Vocab
PARTALO,M., SIJERĈIĆ,Z., 1989.	Examines a selection of speech signal features such as bandpass filter outputs, LPC coefficients and short-time spectrum.	LPC 87.4%, filter bank 83.5%, time domain features 73.8%.	Serbo-Croat digits.
KAMM,C.A., STREETER,L.A., KANE-ESRIG,Y., BURR,D.J., 1989.	Single vowel recognition comparing neural-networks with distance measure techiniques.	The author's elastic measure gave the lowest error rate of around 2-5.5%. Weighted cepstral Euclidean distance also performed well with error rates of 3-4%. Neural network recognition produced results of around 1% error.	Four representations of 11 vowels.

Table A.3. Recognition systems of the 1980s

Author	Speakers	Features	Discussion
PARTALO,M., SIJERĈIĆ,Z., 1989.	Tested on 109 speakers, trained on 55 speakers.	10 LPC coefficients calculated from speech bandlimited to 4.1kHz, preemphasised, 128 sample rectangular windowed (no-overlap). Filter bank response calculated from a 256-point DFT from which a 25 bandpass filter bank output is simulated. Time domain features such as energy and zero-crossings from 12.8ms analysis frame are tested.	DTW matching.
KAMM,C.A., STREETER,L.A., KANE-ESRIG,Y., BURR,D.J., 1989.	-	Perceptual spectrum: derived on 20ms Hamming windowed frames. Frequency scale transformed to Bark scale. LPC specturm from 10 pole LPC analysis. Covariance method on 20ms Hamming widowed frames. Ten cepstral values calculated from the FFTs of the spectra. Distance measures : Formant ratio, LP-residue, Elastic measure (derived by the authors).	Recognition of 20ms vowel portions.

Appendix B

FEATURE BASED CONFUSION TABLES

This appendix gives recognition results obtained during recognition trials described in Chapter 8. The results are tabulated with respect to each speaker tested for both New Zealand and American speakers.

Table B.1. New Zealand Male, speaker independent confusion table for the vocabulary ZERO to NINE. Feature : RMS. Distance : Euclidean.

Input Word	Output Word 0	1	2	3	4	5	6	7	8	9
0	6				1	2				3
1	5	3			2					2
2	3		1			5			2	1
3	5			1	1	2				3
4	1			1	3	3		1	1	2
5	2				2	6			1	1
6	3				2	4	2		1	
7	1	1			3	2		2	2	1
8	2			1	2	1		1	4	1
9	4	1		2	1	3				1

Table B.2. New Zealand Male, speaker independent confusion table for the vocabulary ZERO to NINE. Feature : CEP. Distance : Euclidean.

Input Word	Output Word 0	1	2	3	4	5	6	7	8	9
0	6	1						3		1
1		10			1	1				
2			9	1				2		
3		1		6					4	1
4					11					1
5		1				7				4
6							9	1	2	
7								11	1	
8									12	
9						1				11

Table B.3. New Zealand Male, speaker independent confusion table for the vocabulary ZERO to NINE. Feature : CEP. Distance : Squared.

Input Word	Output Word 0	1	2	3	4	5	6	7	8	9
0	9		1			1				1
1		10				2				
2	1		8	1				1	1	
3			1	8					3	
4					10					2
5	1				2	7				2
6							10	1	1	
7	2							10		
8									12	
9		2								10

Table B.4. New Zealand Male, speaker independent confusion table for the vocabulary ZERO to NINE. Feature : ZCEP. Distance : Euclidean.

Input Word	Output Word 0	1	2	3	4	5	6	7	8	9
0	4	1				2		4		1
1		5			1	2				4
2			6	1				3	2	
3				5		1			6	
4					10					2
5		1				6			2	3
6							11		1	
7								12		
8					1		1		10	
9			1						1	10

Table B.5. New Zealand Male, speaker independent confusion table for the vocabulary ZERO to NINE. Feature : FCEP. Distance : Euclidean.

Input Word	Output Word 0	1	2	3	4	5	6	7	8	9
0	4		1			3		4		
1		5			1	2				4
2			8	1				2	1	
3				6		1		3	2	
4					10					2
5		1				4		1	3	3
6							12			
7								12		
8			1	1					10	
9		1	1			1		4	1	4

Table B.6. New Zealand Female, speaker independent confusion table for the vocabulary ZERO to NINE. Feature : RMS. Distance : Euclidean

Input Word	Output Word 0	1	2	3	4	5	6	7	8	9
0	2	2		2	2	1			2	1
1		6		1	2	1			1	1
2	1		2	1		3			1	4
3	1	1	2	2	1	2				3
4	1	1			4	3				3
5	2	1			1	5			1	2
6				3			7	1	1	
7		1		2		3	1	3	2	
8	1				3	1			7	
9	1	6	1			2		1		1

Table B.7. New Zealand Female, speaker independent confusion table for the vocabulary ZERO to NINE. Feature : CEP. Distance : Euclidean.

Input Word	Output Word 0	1	2	3	4	5	6	7	8	9
0	10	1								1
1		12								
2			11	1						
3			1	11						
4					11	1				
5		1				9				2
6				1			9		2	
7	1		1					10		
8				1	1				10	
9										12

Table B.8. New Zealand Female, speaker independent confusion table for the vocabulary ZERO to NINE. Feature : CEP. Distance : Squared.

Input Word	Output Word 0	1	2	3	4	5	6	7	8	9
0	11	1								
1		12								
2	2		10							
3				11					1	
4					11					1
5	1					10				1
6				3		1	7		1	
7	1							8		3
8					1				11	
9										12

Table B.9. New Zealand Female, speaker independent confusion table for the vocabulary ZERO to NINE. Feature : ZCEP. Distance : Euclidean.

Input Word	Output Word 0	1	2	3	4	5	6	7	8	9
0	8	1								3
1		12								
2		2	3	1				4	2	
3				11					1	
4					11	1				
5		2				8				2
6				2			8		2	
7		1						11		
8					1		1		10	
9										12

Table B.10. New Zealand Female, speaker independent confusion table for the vocabulary ZERO to NINE. Feature : FCEP. Distance : Euclidean.

Input Word	Output Word 0	1	2	3	4	5	6	7	8	9
0	8	1						1		2
1		11								1
2		3	4					3	2	
3			2	8				1	1	
4					11	1				
5	2	3				4				3
6				2			7	1	2	
7		1	2					9		
8				1	1		1		9	
9	1									11

Table B.11. American Male, speaker independent confusion table for the vocabulary ZERO to NINE. Feature : RMS. Distance : Euclidean.

Input Word	Output Word 0	1	2	3	4	5	6	7	8	9
0	5	1	1	2	2		1			
1	2	4	1	4	1					
2			3	6	1				2	
3	2		1	7	1					1
4	1	3	3	1	1		1		2	
5	1	2			6	3				
6		1		2	1		6	1	1	
7		3	1	4	2			2		
8		1	2	4					5	
9	2	3	1	3	1	1	1			

Table B.12. American Male, speaker independent confusion table for the vocabulary ZERO to NINE. Feature : CEP. Distance : Euclidean.

Input Word	Output Word 0	1	2	3	4	5	6	7	8	9
0	11				1					
1		11			1					
2		5	4	1	2					
3		2	1	7						2
4				1	11					
5		4				8				
6							11		1	
7		1	1					10		
8			1	2				1	8	
9		2								10

Table B.13. American Male, speaker independent confusion table for the vocabulary ZERO to NINE. Feature : CEP. Distance : Squared.

Input Word	Output Word 0	1	2	3	4	5	6	7	8	9
0	10	1								1
1		8	1		1					2
2			8		1			2	1	
3		1	1	5		1		1		3
4		2			10					
5						12				
6							11		1	
7	2							10		
8			2				2		8	
9						2				10

Table B.14. American Male, speaker independent confusion table for the vocabulary ZERO to NINE. Feature: ZCEP. Distance : Euclidean.

Input Word	Output Word 0	1	2	3	4	5	6	7	8	9
0	7	5								
1		10		2						
2	1	3	5	2					1	
3		1	2	8						1
4		2			10					
5		1				11				
6							11		1	
7	1	1		1	1			7		1
8			2				1		9	
9		4								8

Table B.15. American Male, speaker independent confusion table for the vocabulary ZERO to NINE. Feature: FCEP. Distance : Euclidean.

Input Word	Output Word 0	1	2	3	4	5	6	7	8	9
0	7	4						1		
1		9			1					2
2		1	7	2	2					
3			2	8					1	1
4		2			9					1
5		3				8				1
6							12			
7		2	1					7		2
8			2						10	
9	5									7

Table B.16. American Female, speaker independent confusion table for the vocabulary ZERO to NINE. Feature : RMS. Distance :Euclidean.

Input Word	Output Word 0	1	2	3	4	5	6	7	8	9
0	4	3			2	2				1
1	3	5			4					
2	1	5		1	2	1		1		1
3		10		1	1					
4	1	1			6	2				2
5	1	4	1		1	1		2		2
6				1	2		7	2		
7	3	3			1	1	1	3		
8		3			3	2			3	1
9		4	1			2				5

Table B.17. American Female, speaker independent confusion table for the vocabulary ZERO to NINE. Feature : CEP. Distance : Euclidean.

Input Word	Output Word 0	1	2	3	4	5	6	7	8	9
0	8		1	1				2		
1	1	9			1					1
2			9	1	1			1		
3		2	1	6					1	2
4		1			11					
5		1			1	10				
6							12			
7			1					11		
8		1	6						4	1
9		2								10

Table B.18. American Female, speaker independent confusion table for the vocabulary ZERO to NINE. Feature : CEP. Distance : Squared.

Input Word	Output Word 0	1	2	3	4	5	6	7	8	9
0	8		2					2		
1		6		1	1			1		3
2		1	8					3		
3		1	3	7		1				
4					12					
5	1				1	11				
6							12			
7			2					10		
8							5		7	
9	1	2								9

Table B.19. American Female, speaker independent confusion table for the vocabulary ZERO to NINE. Feature : ZCEP. Distance : Euclidean.

Input Word	Output Word 0	1	2	3	4	5	6	7	8	9
0	9		3							
1	2	6			1					3
2	1		8	2				1		
3		2		8					1	1
4					12					
5					2	9		1		
6							12			
7			2					10		
8		3	3				1		5	
9		5								7

Table B.20. American Female, speaker independent confusion table for the vocabulary ZERO to NINE. Feature : FCEP. Distance : Euclidean.

Input Word	Output Word 0	1	2	3	4	5	6	7	8	9
0	9		2					1		
1	1	8		1				1		1
2		3	7	1				1		
3		4		6				1	1	
4		1			9			2		
5					2	9		1		
6							12			
7			3					9		
8		3	2				1		6	
9	2	6								4

Table B.21. New Zealand Male, JK, speaker dependent confusion table for the vocabulary ZERO to NINE. Feature : RMS. Distance : Euclidean.

Input Word	Output Word 0	1	2	3	4	5	6	7	8	9
0	6	2	1							3
1		8		1					1	2
2	1		8	2						1
3			4	4			2			2
4	4				5	2		1		
5		4	3		4					1
6		1				2	9			
7	1							11		
8									11	1
9	3	3	1							5

Table B.22. New Zealand Male, JK, speaker dependent confusion table for the vocabulary ZERO to NINE. Feature : CEP. Distance : Euclidean.

Input Word	Output Word 0	1	2	3	4	5	6	7	8	9
0	10		1					1		
1		12								
2			12							
3				12						
4					12					
5		7				1				4
6	1						11			
7								11		1
8									12	
9	1									11

Table B.23. New Zealand Male, JK, speaker dependent confusion table for the vocabulary ZERO to NINE. Feature : CEP. Distance : Squared.

Input Word	Output Word 0	1	2	3	4	5	6	7	8	9
0	10		1					1		
1		12								
2			12							
3				12						
4					11	1				
5		7								5
6	1						10	1		
7								11		1
8									12	
9		1								11

Table B.24. New Zealand Male, JK, speaker dependent confusion table for the vocabulary ZERO to NINE. Feature : ZCEP. Distance : Euclidean.

Input Word \ Output Word	0	1	2	3	4	5	6	7	8	9
0	7		3					1		1
1		12								
2			12							
3				12						
4					10	2				
5		9				1				2
6	1						11			
7								12		
8									12	
9		2							1	9

Table B.25. New Zealand Male, JK, speaker dependent confusion table for the vocabulary ZERO to NINE. Feature : FCEP. Distance : Euclidean.

Input Word \ Output Word	0	1	2	3	4	5	6	7	8	9
0	7		2					1	1	1
1		9				1				2
2			11						1	
3				12						
4					5	7				
5		5				1				6
6	1						11			
7		1						11		
8		1							11	
9		2							1	9

Table B.26. New Zealand Female, TC, speaker dependent confusion table for the vocabulary ZERO to NINE. Feature : RMS. Distance : Euclidean.

Input Word \ Output Word	0	1	2	3	4	5	6	7	8	9
0	6				1	3				2
1		4		1		2				5
2			11							1
3	2		1	6						3
4	2			2	4	3			1	
5	2		3	3	2	2				
6		1		1			7		2	1
7						1		10		1
8								1	11	
9	1	2	1							8

Table B.27. New Zealand Female, TC, speaker dependent confusion table for the vocabulary ZERO to NINE. Feature : CEP. Distance : Euclidean.

Input Word \ Output Word	0	1	2	3	4	5	6	7	8	9
0	8							2		2
1		9				1		1		1
2			12							
3				12						
4		1			11					
5						12				
6							12			
7								12		
8									12	
9										12

Table B.28. New Zealand Female, TC, speaker dependent confusion table for the vocabulary ZERO to NINE. Feature : CEP. Distance : Squared.

Input Word	Output Word 0	1	2	3	4	5	6	7	8	9
0	11							1		
1		12								
2			12							
3				12						
4		1			11					
5						12				
6			1				11			
7								12		
8									12	
9										12

Table B.29. New Zealand Female, TC, speaker dependent confusion table for the vocabulary ZERO to NINE. Feature : ZCEP. Distance : Euclidean.

Input Word	Output Word 0	1	2	3	4	5	6	7	8	9
0	6					1	3	1		1
1		9								3
2			12							
3				11					1	
4					12					
5						12				
6							12			
7							2	10		
8									12	
9										12

Table B.30. New Zealand Female, TC, speaker dependent confusion table for the vocabulary ZERO to NINE. Feature : FCEP. Distance : Euclidean.

Input Word	Output Word 0	1	2	3	4	5	6	7	8	9
0	4	1				2	3	1		1
1	1	5		1			1	1		3
2			9					3		
3				12						
4					12					
5						9	3			
6							12			
7								12		
8									12	
9										12

REFERENCES

AINSWORTH, W. (1967), 'Relative intelligibility of different transforms of clipped speech', *Journal of the Acoustical Society of America*, Vol. 41, No. 5, Pp. 1272–1276.

ALLEN, J. (1985), 'Cochlear modelling', *IEEE Acoustics, Speech and Signal Processing Society Magazine*, Pp. 3–29.

Antex Electronics Corporation (1990), 'Series 2/ Model SX10 digital audio processor datasheet'.

ATAL, B.S. and HANAUER, S. (1971), 'Speech analysis and synthesis by linear prediction', *Journal of the Acoustical Society of America*, Vol. 50, No. 2 (part2), Pp. 637–655.

ATAL, B.S. and SCHOEDER, M. (1986), 'Predictive coding of speech signals', In *The 6th International Congress on Acoustics*, Tokyo, Japan, August 21-28, Pp. 43–48.

AVERBUCH, A., BAHL, L., BAKIS, R., BROWN, P., COLE, A., DAGGETT, G., DAS, S., DAVIS, K., GENNARO, S.D., DE SOUZA, P., EPSTEIN, E., FRALEIGH, D., JELINEK, F., KATZ, S., LEWIS, B., MERCER, R., NADAS, A., NAHAMOO, D., PICHENY, M., SHICHMAN, G. and SPINELLI, P. (1986), 'An IBM-PC based large vocabulary isolated utterance speech recognizer', *International Conference on Acoustics, Speech, and Signal Processing*, April, Pp. 53–56.

BAKER, J.K. (1974), 'The DRAGON system - an overview', *IEEE Symposium on Speech Recognition*, April 15-19, Pp. 22–26.

BAKER, J.M. and PINTO, D.F. (1986), 'Optimal and suboptimal training strategies for automatic speech recognition in noise, and the effects of adaptation on performance', *International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1, No. 14.13, Pp. 745–748. Tokyo.

BARNETT, J. (1973), 'A vocal data management system', *IEEE Transactions on Audio and Electroacoustics*, Vol. AU-21, No. 3, June, Pp. 185–188.

BAUDRY, M. and DUPEYRAT, B. (1982), 'A simple and efficient isolated words recognition system', *International Conference on Acoustics, Speech, and Signal Processing*, Pp. 879–882.

BAUM, L. and EGON, J.A. (1967), 'An inequality with applications to statistical estimation for probabilitic functions of a Markov process and to a model for ecology', *Bull. Amer. Meteorol. Soc.*, Vol. 73, Pp. 360–363.

BAUM, L. and SELL, G.R. (1968), 'Growth functions for transformations on manifolds.', *Pac. J. Mat*, Vol. 27, Pp. 211–227.

BEEK, B., NEUBERG, E.P. and HODGE, D.C. (1977), 'An assessment of the technology of automatic speech recognition for military applications', *IEEE Transactions in Acoustics, Speech and Signal Processing*, Vol. ASSP-25, No. 4, August, Pp. 310–322.

BELLMAN, R. (1957), *Dynamic Programming*, Princeton Univ Press, Princeton, NJ.

BENINGHOF, W.J. and ROSS, M.J. (1970), 'Investigation of an efficient representation of speech spectra for segmentation and classification of speech sounds', *IEEE Transactions on Audio and Electroacoustics*, Vol. AU-18, No. 1, March, Pp. 33–42.

BEZDEL, W. and BRIDLE, J. (1969), 'Speech recognition using zero-crossing meausrements and sequence information', *The Institute of Electrical Engineers*, Vol. 116, No. 4, April, Pp. 617–623.

BEZDEL, W. and CHANDLER, H. (1965), 'Results of an analysis and recognition of vowels by computer using zero-crossing data', *Proceedings of the IEEE*, Vol. 112, No. 11, November, Pp. 2060–2066.

BISIANI, R. and WAIBEL, A. (1982), 'Performance tradeoffs in search techniques for isolated word speech recognition', *International Conference on Acoustics, Speech, and Signal Processing*, Pp. 570–573.

BLADON, A. (1985), *Computer speech processing*, Prentice/Hall International.

BLOOM, P. (1984), 'Use of dynamic programming for automatic synchronization of two similar speech signals', *International Conference on Acoustics, Speech, and Signal Processing*, Pp. 2.6.1–2.6.4.

BRIDLE, J.S., BROWN, M.D. and CHAMBERLAIN, R.M. (1982), 'An algorithm for connected word recognition', *International Conference on Acoustics, Speech, and Signal Processing*, Pp. 899–902.

BROWN, M. and RABINER, L. (1982), 'An adaptive, ordered, graph search technique for DTW isolated word recognition', *IEEE Transactions in Acoustics, Speech and Signal Processing*, Vol. ASSP-30, No. 4, August, Pp. 535–544.

BUSH, M. and KOPEC, G. (1985a), 'Evaluation of a network-based isolated digit recognizer using the TI multi-dialect database', *International Conference on Acoustics, Speech, and Signal Processing*, Vol. 2, No. 23.2.1, Pp. 846–849.

BUSH, M. and KOPEC, G. (1985b), 'Network-based connected digit recognition using vector quantization', *International Conference on Acoustics, Speech, and Signal Processing*, Vol. 3, No. 31.1.1, Pp. 1197–1200.

BUSH, M. and KOPEC, G. (1987), 'Network based connected digit recognition', *IEEE Transactions in Acoustics, Speech and Signal Processing*, Vol. ASSP-35, No. 10, October, Pp. 1401–1412.

CARPENTER, B.E. and LAVINGTON, S.H. (1973), 'The influence of human factors on the performance of a real-time speech recognition system', *Journal of the Acoustical Society of America*, Vol. 53, No. 1, Pp. 42–44.

CHAMBERLAIN, R.M. and BRIDLE, J.S. (1983), 'Zip: a dynamic programming algorithm for time-aligning two indefinitely long utterances', *International Conference on Acoustics, Speech, and Signal Processing*, Pp. 816–819.

CHANDRA, S. and LIN, W. (1974), 'Experimental comparison between stationary and non-stationary formulation of linear prediction applied to voiced speech.', *IEEE Transactions in Acoustics, Speech and Signal Processing*, Pp. 403–401.

CHANG, S.H., PIHL, G.E. and ESSIGMANN, M.W. (1951), 'Representation of speech sounds and some of their statistical properties', *IRE Transactions on Information Theory*, Pp. 147–153.

CHOW, Y.L. and ROUKOS, S. (1989), 'Speech understanding usng a unification grammer', *International Conference on Acoustics, Speech, and Signal Processing*, Vol. 2, No. s13.9, Pp. 727–730.

CHRISTIANSEN, R.W. and RUSHFORTH, C.K. (1977), 'Detecting and locating key words in continuous speech using linear predictive coding', *IEEE Transactions in Acoustics, Speech and Signal Processing*, Vol. ASSP-25, No. 5, October, Pp. 361–367.

CLAPPER, G. (1971), 'Automatic word recognition', *IEEE Spectrum*, August, Pp. 57–69.

CLARK, M.T. (1970), 'Word recognition by means of orthogonal functions', *IEEE Transactions on Audio and Electroacoustics*, Vol. AU-18, No. 3, September, Pp. 304–312.

COLLA, A., SCAGLIOLA, C. and SCIARRA, D. (1985), 'A connected speech recognition system using a diphone-based language model', *International Conference on Acoustics, Speech, and Signal Processing*, Vol. 3, Pp. 1229–1232.

CONDICK, N. and CHALMERS, D. (1989), 'A transputer based speech recognition system', *International Conference on Acoustics, Speech, and Signal Processing*, Vol. 2, No. S15.7, Pp. 797–800.

COOLEY, W.W. and LOHNES, P.R. (1971), *Multivariate data analysis*, John Wiley & Sons, Inc, New York.

CRYSTAL, D. (1980), *Introduction to language pathology*, University park press, Baltimore, 1 ed.

DAS, S.K. (1982), 'Some experiments in discrete utterance recognition', *IEEE Transactions in Acoustics, Speech and Signal Processing*, Vol. ASSP-30, No. 5, October, Pp. 766–782.

DAUTRICH, B., RABINER, L. and MARTIN, T. (1983), 'On the use of filter bank features for isolated word recognition', *International Conference on Acoustics, Speech, and Signal Processing*, Pp. 1061–1064.

DAVIS, S. and MERMELSTEIN, P. (1980), 'Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences', *IEEE Transactions in Acoustics, Speech and Signal Processing*, Vol. ASSP-28, No. 4, August, P. 357.

DAVIS, K., BIDDULPH, R. and BALASHEK, S. (1952), 'Automatic recognition of spoken digits', *Journal of the Acoustical Society of America*, Vol. 24, No. 6, November, Pp. 637–642.

DEMORI, R. (1973), 'A descriptive techinique for automatic speech recognition', *IEEE Transactions on Audio and Electroacoustics*, Vol. AU-21, No. 2, April, Pp. 89–100.

DENES, P. (1959), 'The design and operation of the mechanical speech recognizer at University College London', *Journal British Institute of Radio Engineers(IRE)*, Vol. 19, No. 4, April, Pp. 219–229.

DENES, P. (1964), 'On the motor theory of speech perception', *Models for the perception of speech and visual form*, November, Pp. 309–319. Proceedings of a symposium sponsered by Data Science Laboratory Air Force Cambridge Research Laboratories, Boston, Massachusetts.

DENES, P. and MATHEWS, M. (1960), 'Spoken digit recognition using time-frequency pattern matching', *Journal of the Acoustical Society of America*, Vol. 32, No. 11, November, Pp. 1450–1455.

DREYFUS-GRAF, J. (1950), 'Sonograph and sound mechanics', *Journal of the Acoustical Society of America*, Vol. 22, No. 6, November, Pp. 731–739.

DUDLEY, H. (1939), 'Remaking speech', *Journal of the Acoustical Society of America*, Vol. 11, October, Pp. 169–177.

DUDLEY, H. and BALASHEK, S. (1958), 'Automatic recognition of phonectic patterns in speech', *Journal of the Acoustical Society of America*, Vol. 30, No. 8, August, Pp. 721–732.

DUDLEY, H. and GRUENZ (1946), 'Visible speech translator with external phosphors', *Journal of the Acoustical Society of America*, Vol. 18, No. 1, July.

DUNN, H. (1950), 'The calculation of vowel resonances, and an electrical vocal tract', *Journal of the Acoustical Society of America*, Vol. 22, No. 6, November, Pp. 740–753.

ELDER, A.G. (1991), *Evaluation of glottal characteristics for speaker identification*, University of Canterbury.

ELDER, A., BATES, R., BRIESEMAN, N., CLARK, T., FRIGHT, W., GARDEN, K., KENNEDY, W., SQUIRES, P., TURNER, S. and THORPE, C. (1987), 'Real-time speech therapy aid', *Proceedings of the 24th National Electronics Conference*, September, Pp. 115–118.

ELGHONEMY, M., FIKRI, M., HASHISH, M. and TALKHAN, D. (1986), 'Speaker independent isolated Arabic word recognition system', *International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1, No. 14.1.1, Pp. 697–699.

ERMAN, L.D. (1977), 'A functional description of the Hearsay-II speech understanding system', *International Conference on Acoustics, Speech, and Signal Processing*, May 9-11, Connecticut, Pp. 799–802.

EVERITT (Ed.) (1980), *Clustering Technique*, -.

EWING, G. and TAYLOR, J.F. (1969), 'Computer recognition of speech using zero-crossing information', *IEEE Transactions on Audio and Electroacoustics*, Vol. AU-17, No. 1, March, Pp. 37–40.

FALLSIDE, F. and WOODS, W.A. (Eds.) (1983), *Computer Speech Processing*, Prentice/Hall International.

FANT, G. (1960), *Acoustic Theory of speech production*, Mouton, The Hague.

FRY, D. and DENES, P. (1957), 'On presenting the output of a mechanical speech recognizer', *Journal of the Acoustical Society of America*, Vol. 29, Pp. 364–367.

FRY, D. and DENES, P. (1958), 'The solution of some fundamental problems in mechanical speech recognition', *Language and Speech*, Vol. 1, Pp. 35–58.

FUJIMURA, O. (1975), 'Syllable as a unit of speech recognition', *IEEE Transactions in Acoustics, Speech and Signal Processing*, Vol. ASSP-23, No. 1, February, Pp. 82–87.

FURUI, S. (1981), 'Comparison of speaker recognition methods using statistical features and dynamic features', *IEEE Transactions in Acoustics, Speech and Signal Processing*, Vol. ASSP-29, No. 3, June, Pp. 342–350.

FURUI, S. (1986), 'Speaker-independent isolated word recognition using dynamic features of speech spectrum', *IEEE Transactions in Acoustics, Speech and Signal Processing*, Vol. ASSP-34, No. 1, February, Pp. 52–59.

FURUI, S. (1988), 'A VQ-basd pre-processor using cepstral dynamic features for speaker-independent large vocabulary word recognitoin', *IEEE Transactions in Acoustics, Speech and Signal Processing*, July, Pp. 980–987.

FURUI, S. (1989), 'Unsupervised speaker adaptation method based on hierarchical spectral clustering', *International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1, No. s6.9, Pp. 286–289.

GAGNOULET, C., JOUVET, D. and DAMAY, J. (1991), 'Mairievox: A voice-activated information system', *Speech Communication*, Vol. 210, Pp. 23–31.

GANESAN, K., MARLOT, M. and MEHTA, P. (1986), 'An efficient algorithm for combining vector quanization and stochastic modeling for speaker-independent speech recognition', *International Conference on Acoustics, Speech, and Signal Processing*, Vol. 2, No. 21.2.1, Pp. 1069–1071.

GAUVAIN, J. (1986), 'A syllable-based isolated word recognition experiment', *International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1, No. 2.5.1, Pp. 57–60.

GHITZA, O. (1988), 'Temporal non-place information in the auditory-nerve firing patterns as a front-end for speech recognition in a noisy environment', *Journal of phonetics*, Vol. 16, Pp. 109–123.

GILLI, L. and MEO, A. (1967/68), 'Sequential system for recognizing spoken digits in real time', *Acustica*, Vol. 19, No. 4, Pp. 38–48.

GLASSMAN, M. (1985), 'Hierarchical DP for word recognition', *International Conference on Acoustics, Speech, and Signal Processing*, Pp. 886–889.

GOBL, C. (1989), 'A preliminary study of acoustic voice quality correlates', *STL-QPSR*, Vol. 4/1989, Pp. 9–22.

GODIN, C. and LOCKWOOD, P. (1989), 'DTW schemes for continuous speech recogoniton: a unified view', *Computer Speech and Language*, Vol. 3, Pp. 169–198.

GOLD, B. and RADER, C.M. (1969), *Digital processing of signals*, McGraw-Hill, Inc, New York.

GRAMSS, T. and STRUBE, H.W. (1990), 'Recognition of isolated words based on psychoacoustics and neurobiology', *Speech Communication*, Vol. 9, Pp. 35–40.

GRAY, R.M., BUZO, A., GRAY, A.H. and MATSUYAMA, Y. (1980), 'Distortion measures for speech processing', *IEEE Transactions in Acoustics, Speech and Signal Processing*, Vol. ASSP-28, No. 4, August, Pp. 367–376.

Gray, Jr., A.H. and MARKEL, J.D. (1974), 'A spectral flatness measure for studying the autocorrelation method of linear prediction of speech analysis', *IEEE Transactions in Acoustics, Speech and Signal Processing*, Vol. ASSP-22, December, Pp. 207–217.

Gray, Jr., A.H. and MARKEL, J.D. (1976), 'Distance measures for speech processing', *IEEE Transactions in Acoustics, Speech and Signal Processing*, Vol. ASSP-20, No. 5, October, Pp. 380–391.

HALLE, M. and STEVENS, K. (1962), 'Speech recognition- a model and a program for research', *IRE Transactions on Information Theory*, Vol. IT-8, No. 2, February, Pp. 155–159.

HANSON, B. and WAKITA, H. (1986), 'Spectral slope based distortion measures for all-pole models of speech', *International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1, No. 14.16.1, Pp. 757–760.

HANSON, B.A. and WAKITA, H. (1987), 'Spectral slope distance measures with linear prediction analysis for word recognition in noise', *IEEE Transactions in Acoustics, Speech and Signal Processing*, Vol. ASSP-35, No. 7, July, Pp. 968–973.

HARRIS, F.J. (1978), 'On the use of windows for harmonic analysis with the discrete Fourier transform', *Proceedings of the IEEE*, Vol. 66, No. 1, January, Pp. 51–83.

HATAZAKI, K. and WATANABE, T. (1986), 'A linguistic processor for Japanese continuous speech', *International Conference on Acoustics, Speech, and Signal Processing*, Vol. 2, No. 21.19.1, Pp. 1137–1140.

HATON, J.P. (1974), 'A practical application of a real-time isolated word recognition system using syntactic constraints', *IEEE Transactions in Acoustics, Speech and Signal Processing*, Vol. ASSP-22, No. 5, December, Pp. 416–419.

HERMANSKY, H. (1990), 'Perceptual linear predictive (PLP) analysis of speech', *Journal of the Acoustical Society of America*, Vol. 87, No. 4, April, Pp. 1738–1752.

HERMANSKY, H. and JUNQUA, J.C. (1988), 'Optimization of perceptually-based ASR front-end', *International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1, Pp. 219–222.

HERMANSKY, H., HANSON, B.A. and WAKITA, H. (1985), 'Perceptually based linear predictive analysis of speech', *International Conference on Acoustics, Speech, and Signal Processing*, Pp. 509–512.

HERMANSKY, H., TSUGA, K., MAKINO, S. and WAKITA, H. (1986), 'Perceptually based processing automatic speech recognition', *International Conference on Acoustics, Speech, and Signal Processing*, Pp. 1971–1974.

HILL, D. (1972), 'An abbreviated guide to planning for speech interaction with machines : the state of the art.', *Int. J. Man-Machine studies*, No. 4, Pp. 383–410.

HOHNE, H., COKER, C., LEVINSON, S. and RABINER, L. (1983), 'On temporal alignment of sentences of natural and synthetic speech', *IEEE Transactions in Acoustics, Speech and Signal Processing*, Vol. ASSP-31, No. 4, August, P. 807.

HSU, D. and J.R.DELLER (1989), 'On the use of HMMs to recognize cerebral palsy speech: Isolated word case', *International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1, No. s6.10, Pp. 290–293.

HUNT, M. and LEFEBVRE, C. (1987), 'Speech recognition using an auditory model with pitch synchronous analysis', *International Conference on Acoustics, Speech, and Signal Processing*, Pp. 813–816.

HUNT, M.J. and LEFEBVRE, C. (1989), 'A comparison of several acoustic representations for speech recognition with degraded and undegraded speech', *International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1, No. s6.3, Pp. 262–265.

ICHIKAWA, A., NAKANO, Y. and NAKATA, K. (1973), 'Evaluation of various parameter sets in spoken digit recognition', *IEEE Transactions on Audio and Electroacoustics*, Vol. AU-21, No. 3, June, Pp. 202–209.

IIZUKA, H. (1985), 'Speaker independent telephone speech recognition', *International Conference on Acoustics, Speech, and Signal Processing*, Vol. 2, No. 23.1.1, Pp. 842–845.

ITAHASHI, S. and YOKOYAMA, S. (1976), 'Automatic formant extraction utilizing mel scale and equal loudness contour', *International Conference on Acoustics, Speech, and Signal Processing*, Pp. 310–313.

ITAHASHI, S., MAKINO, S. and KIDO, K. (1973), 'Discrete-word recognition utilizing a word dictionary and phonological rules', *IEEE Transactions on Audio and Electroacoustics*, Vol. AU-21, No. 3, June, Pp. 239–249.

ITAKURA, F. (1974), 'Minimum prediction residual principle applied to speech recognition', *IEEE Symposium on Speech Recognition*, April 15-19, Pp. 101–105.

ITAKURA, F. and SAITO, S. (1968), 'An analysis-synthesis telephony based on maximum likelihood method', *Proc Int Congr Acoust, Tokyo, Japan*, August, Pp. C–5–5.

ITAKURA, F. and SAITO, S. (1970), 'A statistical method for speech spectral density and formant frequencies', *Electronics and Communications in Japan*, Vol. 53-A, No. 1, Pp. 36–42.

JELINEK, F. and OTHERS (1985), 'A real time, isolated-word speech recognition system for dictation transcription', *International Conference on Acoustics, Speech, and Signal Processing*, Pp. 858–861.

JOUVET, D., MONNE, J. and DUBOIS, D. (1986), 'A new network-based speaker-independent connected-word recognition system', *International Conference on Acoustics, Speech, and Signal Processing*, Vol. 2, No. 21.12.1, Pp. 1109–1112.

JUANG, B. (1984), 'On the hidden Markov model and dynamic time warping for speech recognition - a unified view', *AT & T Technical Journal*, Vol. 63, No. 7, September, Pp. 1213–1243.

JUANG, B. and RABINER, L. (1986), 'Mixture autoregressive hidden Markov models for speaker-independent isolated word recognition', *International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1, No. 2.1.1, Pp. 41–44.

JUANG, B., RABINER, L., LEVINSON, S. and SONDHI, M. (1985), 'Recent developments in the application of hidden Markov models to speaker-independent isolated word recognition', *International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1, No. 1.3.1, Pp. 9–12.

JUANG, B.H., RABINER, L.R. and WILPON, J.G. (1987), 'On the use of bandpass liftering in speech recognition', *IEEE Transactions in Acoustics, Speech and Signal Processing*, Vol. ASSP-35, No. 7, July, Pp. 947–954.

JUNQUA, J. (1991), 'A two-pass hybrid system usng low dimensional auditory model for speaker-independent isolated-word recognition', *Speech Communication*, Vol. 10, Pp. 33–44.

JUNQUA, J. and WAKITA, H. (1989), 'A comparative study of cepstral lifters and distance measures for all pole models of speech in noise', *International Conference on Acoustics, Speech, and Signal Processing*, Pp. 476–479.

KAHN, D. and GNANADESIKAN, A. (1986), 'Experiments in speech recognition over the telephone network', *International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1, No. 14.9, Pp. 729–732. Tokyo.

KAMM, C., STREETER, L., KANE-ESRIG, Y. and BURR, D. (1989), 'Comparing performance of spectral distance measures and neural network methods for vowel recognition', *Computer Speech and Language*, Vol. 3, Pp. 21–34.

KAPLAN, H.M. (1971), *Anatomy and physiology of speech*, McGraw-Hill.

KINSLER, L., FREY, A. and COPENS, A. (1982), *Fundamentals of Acoustics*, John Wiley and Sons, 3 ed.

KIRKLAND, J.F. (1993), *Speech Recognition*, University of Canterbury.

KIRKLAND, J. (To be published), -, PhD thesis, Electrical and Electronic Engineering Department, University of Canterbury, Christchurch, New Zealand.

KOENIG, DUNN and LACY (1946), 'The sound spectrograph', *Journal of the Acoustical Society of America*, Vol. 18, July, Pp. 19–49.

KOPP and GREEN (1946), 'Basic phonetic principles of visible speech', *Journal of the Acoustical Society of America*, Vol. 18, July, Pp. 74–89.

LADEFOGAD, P. (1973), *Preliminaries to Linguistic Phonetics.*

LADEFOGAD, P. (1982), *A course in phonetics*, Harcourt Brace Jovanovich, New York, 2 ed.

LAMEL, L.F. and ZUE, V.W. (1982), 'Performance improvement in a dynamic-programming-based isolated word recognition system for the alph-digit task', *International Conference on Acoustics, Speech, and Signal Processing.*

LAU, Y.K. and CHAN, C.K. (1985), 'Speech recognition based on zero crossing rate and energy', *IEEE Transactions in Acoustics, Speech and Signal Processing*, Vol. ASSP-33, No. 1, February, Pp. 320–323.

LAVINGTON (1969), 'Computer simulation of a speech recognition system', *Proceedings of the IEEE*, Vol. 116, No. 6, June, Pp. 1053–1059.

LEA, W.A. (1973), 'An approach to syntactic recognition without phonemics', *IEEE Transactions on Audio and Electroacoustics*, Vol. AU-21, No. 3, June, Pp. 249–258.

LEA, W., MEDRESS, M. and SKINNER, T. (1974), 'A prosodically guided speech understanding strategy', *IEEE Symposium on Speech Recognition*, April 15-19, Pp. 38–44.

LEA, W., MEDRESS, M. and SKINNER, T. (1975), 'A prosodically guided seech understanding strategy', *IEEE Transactions in Acoustics, Speech and Signal Processing*, Vol. ASSP-23, No. 1, February.

LEE, K. (1988), 'Large-vocabulary speaker-independent continuous speech recognition: the SPHINX system', *PhD Disertation, Computer Science Dept, Carnegie Mellon University.*

LEE, C., RABINER, L., PIERACCINI, R. and WILPON, J. (1990), 'Acoustic modeling for large vocabulary speech recognition', *Computer Speech and Language*, Vol. 4, Pp. 127–165.

LESSER, V.R., FENNELL, R.D., ERMAN, L.D. and REDDY, D.D. (1974), 'Organization of the HearsayII speech understanding system', *IEEE Symposium on speech recognition*, April 15-19, Pp. 11–21.

LEVELT, W.J. (1989), *Speaking from intention to articulation*, MIT Press, Canbridge, Massachusetts.

LEVINSON, S.E. and SCHMIDT, C.E. (1983), 'Adaptive computation of articulatory parameters from the speech signal', *Journal of the Acoustical Society of America*, Vol. 74, No. 4, October, Pp. 1145–1154.

LICKLIDER, J. and POLLACK, I. (1948), 'Effects of differentiation, integration and infinite peak clipping upon the intelligibility of speech', *Journal of the Acoustical Society of America*, Vol. 20, No. 1, January, Pp. 42–51.

LIEBERMAN, P. and BLUMSTEIN, S. (1988), *Speech physiology, speech perception and acoustic phonetics*, Cambridge Studies in speech science and communication, Cambridge University.

LINDBLOM, B.E. and SVENSSON, S.G. (1973), 'Interaction between segmental and nonsegmental factors in speech recognition', *IEEE Transactions on Audio and Electroacoustics*, Vol. AU-21, No. 6, December, Pp. 536–545.

LINDGREN, N. (1965a), 'Machine recognition of human language', *IEEE Spectrum*, Vol. 2, No. 3, March, Pp. 114–136. Part I-Automatic speech recognition.

LINDGREN, N. (1965b), 'Machine recognition of human language', *IEEE Spectrum*, Vol. 2, No. 44, April, Pp. 44–59. Part II-Theorectical models of speech perception and language.

LIPPMANN, R. (1987), 'An introduction to computing with neural nets', *IEEE Acoustics, Speech and Signal Processing Society Magazine*, April, Pp. 4–22.

LJOLJE, A. and FALLSIDE, F. (1987), 'Recognition of isolated prosodic patterns using hidden Markov models', *Computer Speech and Language*, Vol. 2, Pp. 27–33.

LOWERRE, B. and REDDY, R. (1980), 'The Harpy speech understanding system', In LEA, W. (Ed.), *Trends in Speech Recogniiton*, Prentice-Hall, Pp. 340–360.

LYON, R. and DYER, L. (1986), 'Experiments with a computational model of the cochlea', *International Conference on Acoustics, Speech, and Signal Processing*, Pp. 1975–1978.

MACLAGAN, M. (1982), 'An acoustic study of New Zealand vowels', *N.Z. Speech Therapist's Journal*, Vol. 37, No. 1, Pp. 20–26.

MANSOUR, D. and JUANG, B.H. (1988), 'A family of distortion measures based upon projection operation for robust speech recognition', *International Conference on Acoustics, Speech, and Signal Processing*, Pp. 36–39.

MARIANI, J. (1989), 'Recent advances in speech processing', *International Conference on Acoustics, Speech, and Signal Processing*, Pp. 429–440.

MARK W. CANNON, J. (1968), 'A method of analysis and recognition for voiced vowels', *IEEE Transactions on Audio and Electroacoustics*, Vol. AU-16, No. 2, June, Pp. 154–158.

MARKEL, J. and A.H. GRAY, J. (1976), *Linear Prediction of Speech*, Spinger-Verlag, Berlin Heidelberg NewYork.

MARRILL, T. (1960), 'Automatic recognition of speech', *IRE Transactions on Human Factors in Electronics*, March, Pp. 34–38.

MEDRESS, M. (1978), 'Speech understanding systems, report of a steering committee', *Artificial Intelligence*, Vol. 9, December, Pp. 307–316.

MEDRESS, M.F., SKINNER, T.E., KLOKER, D.R., DILLER, T.C. and LEA, W.A. (1976), 'A system for the recognition of spoken connected word sequences', *International Conference on Acoustics, Speech, and Signal Processing*, April, Pp. 434–437.

MERCIER, G., CALLEC, A., MONNE, J., QUERRE, M. and TREVARAIN, O. (1982), 'Automatic segmentation, recognition of phonetic units and training in the KEAL speech recognition system', *International Conference on Acoustics, Speech, and Signal Processing*, Pp. 2000–2003.

MERGEL, D. and NEY, H. (1985), 'Phonetically guided clustering for isolated word recognition', *International Conference on Acoustics, Speech, and Signal Processing.*

MERMELSTEIN, P. (1967), 'Determination of the vocal-tract shape from measured formant frequencies', *Journal of the Acoustical Society of America*, Vol. 41, No. 5, Pp. 1281–1293.

MILLER, G.A. (1962), 'Decision units in the perception of speech', *IRE Trans on Info Theory*, Vol. IT-8, February, Pp. 81–83.

MILLER, J., ROSS, P. and WINE, C. (1970), 'An adaptive speech recognition system operating in a remote time-shared computer environment', *IEEE Transactions on Audio and Electroacoustics*, Vol. AU-18, No. 1, March, Pp. 26–31.

MINOUX, M. (1986), *Mathematical programming: Theory and algorithms*, John Wiley and Sons.

MOORE, G.P. (1971), *Organic Voice Disorders*, Prentice-Hall, Inc, Englewood Cliffs, NJ.

MOORE, R.K. (1977), 'Evaluating speech recognizers', *IEEE Transactions in Acoustics, Speech and Signal Processing*, Vol. ASSP-25, No. 2, April, Pp. 178–183.

MOORE, R., RUSSELL, M. and TOMLINSON, M. (1983), 'The discriminating network : a mechanism for focusing recognition in whole-word pattern matching', *International Conference on Acoustics, Speech, and Signal Processing*, Pp. 1041–1044.

MORII, S., NIYADA, K., FUJII, S. and HOSHIMI, M. (1985), 'Large vocabulary speaker-independent Japanese speech recognition system', *International Conference on Acoustics, Speech, and Signal Processing*, Vol. 2, No. 23.7.1, Pp. 866–865.

MOSTELLER, F. (1971), 'The jackknife', *Review of the International sStatistical Institute*, Vol. 39, No. 3, Pp. 363–368.

MYERS, C. and LEVINSON, S. (1982), 'Speaker independent connected word recognition using a syntax-directed dynamic programming technique', *IEEE Transactions in Acoustics, Speech and Signal Processing*, Vol. ASSP-30, No. 4, August, Pp. 561–565.

MYERS, C., RABINER, L. and ROSENBERG, A. (1980), 'Performance tradeoffs in dynamic time warping algorithms for isolated word recognition', *IEEE Transactions in Acoustics, Speech and Signal Processing*, Vol. ASSP-28, No. 6, December, Pp. 623–635.

MYERS, C., RABINER, L. and ROSENBERG, A. (1981), 'On the use of dynamic time warping for word spotting and connected word recognition', *Bell Systems Technical Journal*, Vol. 60, No. 3, March, Pp. 303–325.

NAG, R., AUSTIN, S. and FALLSIDE, F. (1986), 'Using hidden Markov models to define linguistic units', *International Conference on Acoustics, Speech, and Signal Processing*, Vol. 3, Pp. 2239–2242.

NEY, H. (1984), 'The use of a one-stage dynamic programming algorithm for connected word recongition', *IEEE Transactions in Acoustics, Speech and Signal Processing*, Vol. ASSP-32, No. 2, April, Pp. 263–271.

NIEDERJOHN, R.J. (1975), 'A mathematical formulation and comparison of zero-crossing analysis techniques which have been applied to automatic speech recognition', *IEEE Transactions in Acoustics, Speech and Signal Processing*, Vol. ASSP-23, No. 4, August, Pp. 373–379.

NIEDERJOHN, R.J. and THOMAS, I.B. (1973), 'Computer recognition of the continuant phonemes in connected English speech', *IEEE Transactions on Audio and Electroacoustics*, Vol. AU-21, No. 6, December, Pp. 526–535.

NIEDERJOHN, R.J., KRUTZ, M.W. and BROWN, B.M. (1987), 'An experimental investigation of the perceptual effects of altering the zerocrossings of a speech signal', *IEEE Transactions in Acoustics, Speech and Signal Processing*, Vol. ASSP-35, No. 5, May, Pp. 618–625.

NISHIMURA, M. (1989), 'HMM-based speech recognition using dynamic spectral feature', *International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1, No. s6.12, Pp. 298–301.

NOCERINO, N., SOONG, F., RABINER, L. and KLATT, D. (1985), 'Comparative study of several distortion measures for speech recognition', *International Conference on Acoustics, Speech, and Signal Processing*, P. 25.

OLSON, H. and BELAR, H. (1956), 'Phonetic typewriter', *Journal of the Acoustical Society of America*, Vol. 28, No. 6, November, Pp. 1072–1081.

OLSON, H. and BELAR, H. (1960), 'Time compensation for speed of talking in speech recognition machines', *IEEE Transactions on Audio and Electroacoustics*, Vol. AU-8, No. 3, May-June, Pp. 87–90.

OLSON, H. and BELAR, H. (1961), 'Phonectic typewriter III', *Journal of the Acoustical Society of America*, Vol. 33, No. 11, November, Pp. 1610–1615.

OSTENDORF, M. and ROUKOS, S. (1989), 'A stochastic segment model for phoneme-based continuous speech recognition', *IEEE Transactions in Acoustics, Speech and Signal Processing*, Vol. ASSP-37, No. 12, December, Pp. 1857–1869.

PAN, K.C., SOONG, F. and RABINER, L. (1985a), 'A vector quantization based preprocessor for speaker-independent isolated word recognition', *IEEE Transactions in Acoustics, Speech and Signal Processing*, Vol. ASSP-33, No. 3, june, Pp. 546–560.

PAN, K., SOONG, F., RABINER, L. and BERGH, A. (1985b), 'An efficient vector-quantization preprocessor for speaker-independent isolated word recognition', *International Conference on Acoustics, Speech, and Signal Processing*, Vol. 2, No. 23.9.1, Pp. 874–877.

PARTALO, M. and SIJERCIC, Z. (1988), 'Comparison of several speech signal feature parameters for automatic speech recognition', *Speech Communication*, Vol. 8, Pp. 347–353.

PIERCE, J. (1969), 'Whither speech recognition?', *Journal of the Acoustical Society of America*, Vol. 46, No. 4,pt2, October, Pp. 1049–1051.

POLS, L.C. (1971), 'Real-time recognition of spoken words', *IEEE Transactions Computing*, Vol. C-20, September, Pp. 972–978.

POTTER and PETERSON (1948), 'The representation of vowels and their movements', *Journal of the Acoustical Society of America*, Vol. 20, No. 4, July, Pp. 528–535.

POTTER, R. and STEINBERG, J. (1950), 'Toward the specification of speech', *Journal of the Acoustical Society of America*, Vol. 22, No. 6, November, Pp. 807–820.

PREECE and STROH (1879), '–', *Proc Royal Society*, Vol. xxviii, P. 359.

PURTON, R.F. (1968), 'Speech recognition using autocorrelation analysis', *IEEE Transactions on Audio and Electroacoustics*, Vol. AU-16, No. 2, June, Pp. 235–239.

RABINER, L. (1984), 'On the applications of energy contours to the recognition of connected word sequences', *Bell Systems Technical Journal*, Vol. 63, No. 9, November, Pp. 1981–1995.

RABINER, L. (1989), 'A tutorial on hidden Markov models and selected applications in speech recognition', *Proceedings of the IEEE*, Vol. 77, No. 2, February, Pp. 257–285.

RABINER, L.R. and GOLD, B. (1975), *Theory and application of digital signal processing*, Prentice-Hall, Inc.

RABINER, L.R. and LEVINSON, S.E. (1981), 'Isolated and connected word recognition - theory and selected applications', *IEEE Transactions in Communications*, Vol. COM-29, No. 5, May, Pp. 621–659.

RABINER, L.R. and LEVINSON, S.E. (1985), 'A speaker-independent, syntax directed, connected word recognition system based on hidden Markov models and level building', *IEEE Transactions in Acoustics, Speech and Signal Processing*, Vol. ASSP-33, No. 3, June, Pp. 561–573.

RABINER, L. and SAMBUR, M. (1975), 'An algorithm for determining the endpoints of isolated utterances', *Bell Systems Technical Journal*, Vol. 54, No. 2, February, Pp. 297–315.

RABINER, L.R. and SAMBUR, M.R. (1976), 'Some preliminary experiments in the recognition of connected digits', *IEEE Transactions in Acoustics, Speech and Signal Processing*, Vol. ASSP-24, April, Pp. 170–182.

RABINER, L.R. and SCHAFER, R.W. (1978), *Digital Processing of Speech Signals*, Prentice-Hall, Inc.

RABINER, L.R. and SCHMIDT, C.E. (1980), 'Application of dynamic time warping to connected digit recognition', *IEEE Transactions in Acoustics, Speech and Signal Processing*, Vol. ASSP-28, No. 4, August, Pp. 377–388.

RABINER, L.R., ROSENBERG, A.E., WILPON, J.G. and KEILIN, W.J. (1982), 'Isolated word recognition for large vocabularies', *Bell Systems Technical Journal*, Vol. 61, No. 10, December, Pp. 2989–3005.

RABINER, L., LEVINSON, S. and SONDHI, M. (1983), 'On the application of vector quantization and hidden Markov models to speakr-independent, isolated word recognition', *Bell Systems Technical Journal*, Vol. 62, No. 4, April, Pp. 1075–1105.

RABINER, L., SONDHI, M. and LEVINSON, S. (1984a), 'A vector quantizer combining energy and LPC parameters and its application to isolated word recognition', *AT & T Technical Journal*, Vol. 63, No. 5, May, Pp. 721–736.

RABINER, L.R., PAN, K.C. and SOONG, F.K. (1984b), 'On the performance of isolated word speech recognizers using vector quantization and temporal energy contours', *AT & T Technical Journal*, Vol. 63, No. 7, September, Pp. 1245–1260.

RABINER, L., WILPON, J. and JUANG, B. (1986), 'A continuous training procedure for connected digit recognition', *International Conference on Acoustics, Speech, and Signal Processing*, Vol. 2, No. 21.1.1, Pp. 1065–1068.

RABINER, L., WILPON, J. and SOONG, F. (1989), 'High performance connected digit recognition using hidden Markov models', *IEEE Transactions in Acoustics, Speech and Signal Processing*, Vol. ASSP-37, No. 8, August, Pp. 1214–1225.

RAYLEIGH, L. (1878), *The theory of sound*, Vol. 2, –, –.

REDDY, D. (1966), 'Segmentation of speech sounds', *Journal of the Acoustical Society of America*, Vol. 40, No. 2, March, Pp. 307–312.

REDDY, D. (1967a), 'Phoneme grouping for speech recognition', *Journal of the Acoustical Society of America*, Vol. 41, No. 5, September, Pp. 1295–1301.

REDDY, D. (1967b), 'Computer recognition of connected speech', *Journal of the Acoustical Society of America*, Vol. 42, No. 2, Pp. 329–347.

REDDY, D.R. (1969), 'Segment-synchronization problem in speech recognition', *Journal of the Acoustical Society of America*, P. 89(A).

REDDY, D.R. (1976), 'Speech recognition by machine: A review', *Proceedings of the IEEE*, Vol. 64, No. 4, April, Pp. 501–531.

REDDY, D.R., ERMAN, L.D. and NEELY, R.B. (1973), 'A model and a system for machine recognition of speech', *IEEE Transactions Audio Electroacoustics*, Vol. AU-21, June, Pp. 229–238.

RIESZ and SCHOTT (1946), 'Visible speech cathode ray translator', *Journal of the Acoustical Society of America*, Vol. 18, July, Pp. 4–18.

RITEA, H. (1974), 'A voice controlled data management system', *IEEE Symposium on Speech Recognition*, Pp. 28–31.

ROACH, P., ROACH, H., DEW, A. and ROWLANDS, P. (1990), 'Phonetic analysis and the automatic segmentation and labelling of speech sounds.', *Journal of the International phonetic Association*, Vol. 20:1, Pp. 15–21.

ROSENBER, A., RABINER, L., WILPON, J. and KAHN, D. (1983), 'Demisyllable-based isolated word recognition system', *IEEE Transactions in Acoustics, Speech and Signal Processing*, Vol. ASSP-31, No. 3, june, P. 713.

ROSENBERG, A. and ITAKURA, F. (1976), 'Evaluation of an automatic word recognition system over dialed-up telephone lines', *Journal of the Acoustical Society of America*, Vol. 60,suppl 1, P. s12(A).

RUSKE, G. (1982), 'Automatic recognition syllabic speech segments using spectral and temporal features', *International Conference on Acoustics, Speech, and Signal Processing*, Pp. 550–553.

RUSSELL (1929), 'The mechanism of speech', *Journal of the Acoustical Society of America*, Vol. 1, No. 1, Pp. 83–109.

RUSSELL, M. and MOORE, R. (1985), 'Explicit modelling of state occupancy in hidden Markov models for automatic speech recognition', *International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1, No. 1.2.1, Pp. 5–8.

RUSSELL, M., MOORE, R. and TOMLINSON, M. (1983), 'Some techniques for incorporating local timescale variability information into a dynamic time-warping algorithm for automatic speech recognition', *International Conference on Acoustics, Speech, and Signal Processing*, Pp. 1037–1040.

RUSSELL, M., DEACON, J. and MOORE, R. (1984), 'Some implications for the effect of template choice on the performance of an automatic speech recogniser', *Proc I.O.A.*, Vol. 6, No. 4, Pp. 287–292.

SAKAI, T. and DOSHITA, S. (1963), 'The automatic speech recognition system for conversational sound', *IEEE Transactions on Electronic Computers*, December, Pp. 835–846.

SAKOE, H. (1979), 'Two-level DP matching - a dynamic programming based pattern matching alorithm for connected word recognition', *IEEE Transactions in Acoustics, Speech and Signal Processing*, Vol. ASSP-27, No. 6, December, Pp. 588–595.

SAKOE, H. and CHIBA, S. (1970), 'A similarity evaluation of speech patterns by dynamic programming', *Insst Electron Comm Eng Japan*, P. 136. in Japanese.

SAKOE, H. and CHIBA, S. (1978), 'Dynamic programming algorithm optimization for spoken word recognition', *IEEE Transactions in Acoustics, Speech and Signal Processing*, Vol. ASSP-26, No. 1, February, Pp. 43–49.

SAMBUR, M.R. and RABINER, L.R. (1975), 'A speaker-independent digit-recognition system', *The Bell System Technical Journal*, January, Pp. 81–102.

SAMBUR, M.R. and RABINER, L.R. (1976), 'A statistical decision approach to the recognition of connected digits', *IEEE Transactions in Acoustics, Speech and Signal Processing*, Vol. ASSP-24, No. 6, December, Pp. 550–559.

SAUTER, L. (1985), 'Isolated word recognition using a segmental approach', *International Conference on Acoustics, Speech, and Signal Processing*, Vol. 2, No. 23.3.1, Pp. 850–852.

SAVOJI, M. (1989), 'A robust algorithm for accurate endpointing of speech signals', *Speech Communication*, Vol. 8, No. 1, March, Pp. 45–60.

SCARR, R. (1968), 'Zero crossing as a means of obtaining spectral information in speech analysis', *IEEE Transactions on Audio and Electroacoustics*, Vol. AU-16, No. 2, June, Pp. 247–255.

SCARR, W. (1970), 'Word-recognition machine', *Proceedings IEE*, Vol. 117, No. 1, January, Pp. 203–212.

SCHWARTZ, R., CHOW, T., KIMBALL, O., ROUCOS, S., KRASNER, M. and MAKHOUL, J. (1985), 'Context-dependent modelling for acoustic-phonetic recognition of continuous speech', *International Conference on Acoustics, Speech, and Signal Processing*, Vol. 3, No. 21.15.1, Pp. 1121–1125.

SENEFF, S. (1986), 'A computational model for the peripheral auditory system application to speech recognition research', *International Conference on Acoustics, Speech, and Signal Processing*, Pp. 1983–1986.

SHAW, E. (1980), *Acoustical factors affecting hearing and performance*, -.

SHEARME, J. and LEACH, P. (1968), 'Some experiments with a simple word recognition system', *Trans on audio and electroacoustics*, Vol. 16, No. 2, June, Pp. 256–261.

SHORE, J.E. and BURTON, D. (1982), 'Discrete utterance speech recognition without time normalization', *International Conference on Acoustics, Speech, and Signal Processing*, Pp. 907–910.

SKINNER, P.H. and SHELTON, R.L. (1978), *Speech, Language, and hearing: normal processes and disorders*, Addison-weley publishing company, Reading, Massachusetts.

SMITH, C. (1951), 'A phoneme detector', *Journal of the Acoustical Society of America*, Vol. 23, No. 4, July, Pp. 446–451.

S.NAKAGAWA and JILAN, M. (1986), 'Syllable-based connected spoken word recognition by two-pass O(n) DP matching and hidden Markov models', *International Conference on Acoustics, Speech, and Signal Processing*, Vol. 2, No. 21.14.1, Pp. 1117–1120.

SONDHI, M.M. (1979), 'Estimation of vocal-tract areas : The need for acoustical measurements', *IEEE Transactions in Acoustics, Speech and Signal Processing*, Vol. ASSP-27, No. 3, June, Pp. 268–273.

SONDHI, M. and GOPINATH, B. (1971), 'Determination of vocal-tract shape from impulse response at the lips', *Journal of the Acoustical Society of America*, Vol. 49, No. 6, Pp. 1867–1873.

SOONG, F.K. and ROSENBERG, A.E. (1988), 'On the use of instantaneous and transitional spectral information in speaker recognition', *IEEE Transactions in Acoustics, Speech and Signal Processing*, Vol. ASSP-36, No. 6, June, Pp. 871–879.

SRIVASTAVA, M. and KHATRI, C. (1979), *An introduction to Multivariate Statistics*, Elsevier North Holland, Inc, 52 Vanderbilt Avenue, New York, New York.

STEINBERG and FRENCH (1946), 'The portrayal of visible speech', *Journal of the Acoustical Society of America*, Vol. 18, July, Pp. 4–18.

SUBRATA, K. (1982), 'Some experiments in discrete utterance recognition', *IEEE Transactions in Acoustics, Speech and Signal Processing*, Vol. ASSP-30, No. 5, October, P. 766.

SUGAWARA, K., NISHIMURA, M., TOSHIOKA, K., OKOCHI, M. and KANEKO, T. (1985), 'Isolated word recognition using hidden Markov models', *International Conference on Acoustics, Speech, and Signal Processing*, Pp. 1–4.

SVENDSEN, T., PLIWAL, K., HARBORG, E. and HUSOY, P. (1989), 'An improved sub-word based speech recognizer', *International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1, No. s3.8, Pp. 108–111.

TEACHER, C.F., KELLET, H.G. and FOCHT, L.R. (1967), 'Experimental, limitied vocabulary, speech recognizer', *IEEE Transactions on Audio and Electroacoustics*, Vol. AU-15, No. 3, September, Pp. 127–130.

THORPE, C.W. (1990), *Processing of speech and other sounds.*, PhD thesis, Electrical and Electronic Engineering Department, University of Canterbury, Christchurch, New Zealand.

TOGAWA, F., HAKARIDANI, M., IWAHASHI, H. and UEDA, T. (1986), 'Voice activated word processor with automatic learning for dynamic optimization of syllable-templates', *International Conference on Acoustics, Speech, and Signal Processing*, Vol. 2, No. 21.15.1, Pp. 1121–1125.

TOHKURA, Y. (1986), 'A weighted cepstral distance measure for speech recognition', *International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1, No. 14.17.1, Pp. 761–764.

TOHKURA, Y. (1987), 'A weighted cepstral distance measure for speech recognition', *IEEE Transactions in Acoustics, Speech and Signal Processing*, Vol. ASSP-35, No. 10, October, Pp. 1414–1422.

TREHERN, J., JACK, M. and LAVER, J. (1986), 'Speech processing wih a Boltzmann machine', *International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1, No. 14.6.1, Pp. 721–724.

TRIBOLET, J., RABINER, L. and WILPON, J. (1982), 'An improved model for isolated word recognition', *Bell Systems Technical Journal*, Vol. 61, No. 9, November, Pp. 2289–2313.

VELICHIKO, V. and ZAGORUYKO, N. (1970), 'Automatic recognition of 200 words', *Int. J. Man-Machine Studies*, Vol. 2, Pp. 223–234.

VIDAL, E. and LLORET, M.J. (1988), 'Fast speaker independent DTW recognition of isolated words using a metric space search algorithm(AESA)', *Speech Communication*, Vol. 7, Pp. 417–422.

VIDAL, E., RULOR, H.M., CASACUBERTA, F. and BENEDI, J.M. (1988), 'On the use of a metric space search algorithm(AESA) for fast DTW-based recognition of isolated words', *IEEE Transactions in Acoustics, Speech and Signal Processing*, Vol. ASSP-36, No. 5, May, Pp. 651–659.

VONKELLAR, T. (1971), 'An on-line recognition system for spoken digits', *Journal of the Acoustical Society of America*, Vol. 49, No. 4, part 2, Pp. 1288–1296.

WAKITA, H. (1973), 'Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms', *IEEE Transactions on Audio and Electroacoustics*, Vol. AU-21, No. 5, October, Pp. 417–427.

WALKER, D.E. (1974), 'The SRI speech understanding system', *IEEE symposium on speech recognition*, April 15-19, Pp. 32–37.

WARREN, J.H. (1971), 'A pattern classification technique for speech recognition', *IEEE Transactions on Audio and Electroacoustics*, Vol. AU-19, No. 4, December, Pp. 281–285.

WATARI, M., AKABANE, M. and SAKO, Y. (1983), 'A speaker-independent word recognition based on transient matching', *International Conference on Acoustics, Speech, and Signal Processing*, Pp. 715–718.

WHITE, G.M. and NEELY, R.B. (1976), 'Speech recognition experiments with linear prediction, bandpass filtering and dynamic programming', *IEEE Transactions in Acoustics, Speech and Signal Processing*, Vol. ASSP-24, No. 2, April.

WIREN and STUBBS (1956), 'Electronic binary selection system for phoneme classification', *Journal of the Acoustical Society of America*, Vol. 28, No. 6, November, Pp. 1082–1091.

WITTEN, I. (1982), *Principles of computer speech*, Academic press.

WOODS, W.A. (1975), 'Motivation and overview of SPEECHLIS: An experimental prototype for speech understanding research', *IEEE Transactions in Acoustics, Speech and Signal Processing*, Vol. ASSP-23, February, Pp. 2–10.

YATO, F., ASAMIA, T. and HIGUCHI, N. (1986), 'Speech understanding system using knowledge engineering techniques', *International Conference on Acoustics, Speech, and Signal Processing*, Vol. 2, No. 21.18.1, Pp. 1133–1136.

ZELINSKY, R. and NOLL, P. (1977), 'Adaptive transform coding of speech signals', *IEEE Transactions in Acoustics, Speech and Signal Processing*, Vol. ASSP-25, August, Pp. 299–309.